Predictive computational modeling for improved treatment strategies

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium
(Dr. rer. nat.)

im Fach
Biophysik

eingereicht an der
Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Dipl.-Phys. Max Schelker

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät:
Prof. Dr. Bernhard Grimm

Gutachter/innen:

1. Prof. Dr. Dr. h.c. Edda Klipp
2. Prof. Dr. Hanspeter Herzel
3. Prof. Dr. Jens Timmer

Tag der mündlichen Prüfung: 29.08.2017

Max Schelker
Predictive computational modeling for improved treatment strategies
May 2, 2017
Humboldt-Universität zu Berlin
Theoretical Biophysics
Department of Biology
Faculty of Life Sciences
Invalidenstr. 42
10115 Berlin

Abstract

Cancer and infectious diseases, such as influenza infection, represent major threats to the human population, especially since demographic change makes more and more people vulnerable. Mathematical modeling of disease covers several layers of detail ranging from epidemiological models for infection spread to cancer-associated signaling within individual cells. These models, when being calibrated to biological data, can provide useful means for generating hypothesis of priorly unknown interactions, predicting drug targets for novel therapeutic substances and for improving the understanding and efficient functioning of existing treatment strategies. In this thesis, I present several projects in which predictive computational models are utilized to gain deeper insights into the biological processes and to improve therapy of cancer and associated health problems.

The first part highlights the importance of community-driven software development for systems biology applications. Efficient, yet expandable and open software continuously improves, driven by an active community of users and developers.

In the second project, the intracellular processes during the early influenza A infection are investigated. Using a combination of experimental measurements and mathematical modeling, degradation of the viral genome during its diffusion through the cytoplasm could be identified as a limiting factor for a successful infection. By experimentally increasing the pH sensitivity of the viral hemagglutinin protein, the distance of diffusion was increased and the computationally predicted decrease in infectivity could be validated in experiment.

The third project deals with cancer-associated health issues and their treatment. Patients suffering from anemia, caused by the cancer itself or as a side-effect of chemotherapy, are treated either with blood transfusions or with an erythropoiesis stimulating agent (ESA). By adapting a published model of ESA-EpoR interaction, not only the biochemical properties of different ESAs could be characterized *in silico* but also the number of binding sites (i.e. Epo receptors on the cell surface) in different cell lines was accurately determined. The model was extended by a pharmaco-kinetic and -dynamic part. The combined ESA-EpoR-PK/PD model could be utilized to retrospectively optimize the dosing regimen of patients suffering from anemia.

In the last project, a computational method for analyzing and deciphering the cellular composition of bulk tumor samples is presented. Only recently, a new class of anti-cancer drugs was introduced recruiting the body's own immune system to combat malignant

tissue. However, the efficient functioning of these immunotherapeutical drugs heavily depends on the presence of specific immune cells in the tumor micro-environment. Based on single-cell RNA sequencing data, an existing method for computational deconvolution could be adapted for data from solid tumor tissue and its performance was benchmarked.

Taken together, in this thesis I present approaches how predictive computational models can be utilized to render more efficient existing treatment strategies.

Zusammenfassung

Krebs und Infektionskrankheiten, wie z.B. Influenza, stellen zwei der großen Bedrohungen für die Menschheit dar. Gerade durch den demographischen Wandel sind immer mehr Menschen gefährdet. Mathematische Modelle von Krankheiten decken verschiedene Detailebenen ab – von epidemologischen Modellen der Virusinfektion bis zu intrazellulären Modellen der Signaltransduktion in einzelnen Krebszellen. Diese Modelle, sofern sie anhand von biologische Daten kalibriert wurden, können sich als sehr nützlich erweisen um Hypothesen zu bisher unbekannten Wechselwirkungen zu generieren, Wirkstoffkandidaten vorherzusagen, und die Funktionsweise von existierenden Wirkstoffen besser zu verstehen.

Im Rahmen dieser Arbeit möchte ich mehrere Projekte vorstellen, in denen prädiktive mathematische Modelle dazu benutzt wurden, tiefere Einblicke in die biologischen Prozesse zu gewinnen und die Therapieansätze bei Krebserkrankungen und den damit verbundenen Gesundheitsproblemen zu verbessern.

Im ersten Teil geht es um die Bedeutung von gemeinschaftlich entwickelter Software für die Systembiologie. Eine offene und erweiterbare Modellierungssoftware ermöglicht es ständig verbessert zu werden und an die Bedürfnisse der Nutzer angepasst zu werden.

Im zweiten Projekt wurden die intrazellulären Prozesse während der frühen Influenza A Infektion untersucht. Durch eine Kombination von biologischen Messungen und mathematischer Modellierung konnte der Abbau von viraler RNA währernd des Transportes durch das Wirtszellzytoplasma als limitierender Faktor für die erfolgreiche Infektion identifiziert werden. Mit Hilfe eines experimentell modifizierten viralen Hämagglutintin-Proteins mit veränderter pH-Abhängigkeit konnte gezeigt werden, dass sich der Abstand

zum Zellkern, in dem das virale Genom freigesetzt wird, vergrößert. Die Modellvorhersage, dass die Infektion dadurch weniger effektiv wird, konnte experimentell bestätigt werden.

Im dritten Projekt beschäftigte ich mich mit gesundheitlichen Problemen, die im Zusammenhang mit einer Krebserkrankung und deren Behandlung auftreten können. Chemotherapie oder die Krebserkrankung selbst führt bei vielen Patienten zu einer Blutarmut (*Anämie*). Diese wird aktuell entweder durch regelmäßige Bluttransfusionen oder durch Verabreichung von sogenannten *Erythropoiesis-Stimulating Agents* (kurz: ESA, zu Deutsch: Erythropoese-stimulierende Substanzen) behandelt. Mithilfe eines publizierten mathematischen Modells zur ESA-EpoR Interaktion konnten die Bindungseigenschaften verschiedener ESAs charakterisiert und zudem die Anzahl der Bindungsstellen auf unterschiedlichen Zelllinien bestimmt werden. Durch eine Erweiterung des Modells mit einem pharmakokinetischen und -dynamischen Teil konnte die Dosierung für Anämiepatienten retrospektiv verbessert werden.

Das letzte Projekt stellt eine computerbasierte Methode zur Analyse und Entschlüsselung der zellulären Zusammensetzung von Tumorproben dar. In den vergangenen Jahren wurde vermehrt eine neue Klasse von Krebsmedikamenten entwickelt, die sich das körpereigene Immunsystem zunutze macht, um den Krebs zu bekämpfen. Das Funktionieren dieser Medikamente hängt jedoch davon ab, ob bestimmte Immunzellen in der Umgebung des Tumors vorhanden sind. Auf Grundlage von Einzelzell RNA-Sequenzierungsdaten konnte eine existierende Methode so erweitert werden, dass nunmehr auch Proben von soliden Tumoren entschlüsselt werden können. Zudem wurden die Einflüsse von verschiedenen Faktoren, wie etwa der Gewebeherkunft oder dem verwendeten Algorithmus, systematisch ausgewertet.

Zusammengefasst habe ich in dieser Arbeit dargestellt, wie prädiktive Computermodelle dazu verwendet werden können bestehende Behandlungsansätze zu verbessern und neue Wirkstoffkandidaten zu identifizieren.

Acknowledgement

First of all, I want to thank my supervisor **Edda Klipp**. She did not only offer me funding for more than four years, but also gave me the freedom to grow into an independently working researcher and gave me all the opportunities to discover the field of systems biology.

Next, a big “thank you” goes to my colleagues from the group of theoretical biophysics, especially to **Björn Goldenbogen**, **Max Flöttmann**, **Katja Tummler** and **Marcus Krantz**, who joined me on many adventures – from the first attempts of kitesurfing to regular Freeletics workouts in the lunch break to the pleasures of sky diving – and enriched my daily work life with *fruitful* discussions.

Furthermore, I thank **Andreas Raue**, who invited me to join Merrimack Pharmaceuticals in Cambridge, MA, USA where I had an extremely inspiring time and discovered the joys of sailing.

I also want to thank all my collaboration partners from **Berlin**, **Cambridge**, **Heidelberg** and **Freiburg**. With endurance and diligence, they acquired thousands of data points and enabled our models to grow to something bigger.

Last but not least, I thank my dear **family** for always being there, my girlfriend **Lisa** for her continuous loving support over the last years and my baby daughter for giving me some extra time to finish this thesis.

Contents

1 Introduction **1**

2 Methods **5**

2.1 Experimental methods 5

2.1.1 Fluorescence microscopy 5

2.1.2 Flow cytometry and cell sorting 5

2.1.3 Quantitative immunoblotting 6

2.1.4 Enzyme-linked immunosorbent assay (ELISA) 7

2.1.5 Quantitative real-time PCR (qPCR) 7

2.1.6 High-throughput RNA sequencing (RNA-seq) 8

2.2 Data analysis 10

2.2.1 RNA-seq data processing 11

2.2.2 Dimensionality reduction 13

2.2.3 Clustering 14

2.2.4 Support Vector Regression (SVR) 15

2.2.5 Decision-tree Classification 16

2.3 Dynamic modeling 17

2.3.1 Biochemical reaction networks 17

2.3.2 Parameter estimation 19

2.3.3 Parameter identifiability 22

2.4 Numerical methods 24

2.4.1 Integration of differential equations 24

2.4.2 Optimization 25

3 Developing software tools tailored to systems biology applications **31**

3.1 Data2Dynamics – A framework for data-based modeling 31

3.1.1 Community based development 33

3.1.2 Selected features 35

3.1.3 Benchmark of sampling strategies and implemented fitting algorithms 38

3.2 A data interface for the ViroSign project 43

3.3 Conclusions 44

4 Mathematical modeling of the influenza A infection **47**
4.1 Introduction to influenza biology 47
4.1.1 The influenza A virus 47
4.1.2 Epidemiology of influenza A 48
4.1.3 The replication cycle of influenza A 49
4.2 Dynamic modeling of the influenza A infection 52
4.2.1 Existing models 53
4.2.2 An ODE model describing influenza A virus entry into the host cell 54
4.2.3 A stochastic model of vRNP diffusion to the nucleus 59
4.2.4 Combining both modeling approaches reveals vRNP removal in the cytosol 63
4.3 Discussion 64

5 Computational approaches for optimized treatment of cancer-associated health conditions **67**
5.1 Introduction 67
5.2 Optimized treatment strategies for anemia patients based on a mechanistic multi-scale model 69
5.2.1 Introduction 69
5.2.2 Building a mathematical model of ESA-EpoR interaction 71
5.2.3 Extending the model for multiple cell types and ESAs 75
5.2.4 Incorporating ligand depletion experiments for various cell lines and ESAs 77
5.2.5 Results 79
5.2.6 Linking the receptor model to JAK-STAT signaling 88
5.2.7 Pharmacokinetics and -dynamics (PKPD) 90
5.2.8 Discussion 93
5.3 Deciphering the cellular composition of unknown patient samples for immunotherapy 97
5.3.1 Introduction 97
5.3.2 Single-cell RNA-seq data from multiple tissues and patients . . . 101
5.3.3 Classifying individual cells based on marker gene expression and similarity 103
5.3.4 Analyzing indication and patient-specific gene expression 108
5.3.5 Defining a benchmark based on single-cell gene expression data . 110
5.3.6 Impact of signature gene sets on deconvolution accuracy 114
5.3.7 Evaluating deconvolution algorithms 114
5.3.8 Impact of patient- and indication-specific reference profiles 115
5.3.9 Estimating T cell subtypes and prognostic ratios 117

5.3.10 Predicting tumor expression profiles 119
5.3.11 Discussion 121
5.4 Conclusions 123

6 Conclusions & Outlook **125**
6.1 Conclusions 125
6.2 Outlook 130

Bibliography **133**

A Mathematical modeling of the influenza A infection **149**

B Computational approaches for optimized treatment of cancer-associated health conditions **153**
B.1 Optimized treatment strategies for anemia patients based on a mechanistic multi-scale model 153
B.1.1 Auxiliary ESA-EpoR model 153
B.2 Deciphering the cellular composition of unknown patient samples for immunotherapy 161

Declaration **167**

1 Introduction

> *Quite soon, as we move from genes to the proteins that they code for, and then on to the interactions between these proteins, the problems become seriously complicated. [142]*
>
> **— Denis Noble**
> *The Music of Life: Biology beyond the Genome*

When the *Human Genome Project* announced in 2001 that more than $90\,\%$ of the sequence of the human genome was decoded [110], more questions were raised than answered. Despite the big success of having almost all genetic information of a human being available, functional dependencies remained unclear and the role of epigenetic modifications and regulation of gene expression became more evident than ever. Thus, understanding the functional behavior requires more than knowing the genetic code or the amino-acid sequence of a protein. But how can we improve the *bigger picture* when we focus on all the small details? In systems biology, interdependencies in complex systems are disentangled through a combination of experimental approaches and mathematical modeling. Furthermore, recent advances in the field of bioinformatics help to extract valuable information from high-dimensional data obtained by modern high-throughput technologies.

In this thesis, several projects making use of methods from computational systems biology are presented. While their biological background might seem unconnected, all of them rely on *predictive mathematical models* in one or the other way, thereby forming the core of this thesis. Before going more into detail, I want to shortly introduce the projects and motivate their research questions.

Software development for systems biology

In computational systems biology, methods originating from statistics, mathematics and physics are adapted to describe the behavior of biological systems. Complex experimental setups with various readouts and indirect observations need to be mapped to trajectories of differential equation models in order to calibrate priorly unknown parameters. As most

models are non-linear in their parameters, optimization is not trivial and requires efficient and robust implementations of established algorithms and methods. Since the early beginnings of computational systems biology in the late 1990s [98], various software tools have been developed to enable both experimentalist and theoreticians to perform model simulations [83, 178, 124]. However, many of those software solutions are relatively problem-specific or sacrifice computational efficiency for graphical interfaces designed for users without a profound theoretical background. Starting from the great need for curated solutions for data-based computational modeling, a group of researchers developed the *Data2Dynamics* modeling software for MATLAB. One by one, the basic concepts as well as advanced features got implemented by a growing community of developers continuously extending the range of possible use cases. The questions that arise here are:

Can community-driven software development accelerate and improve workflows in computational biology research? What features do software solutions need to facilitate model construction, calibration and analysis for systems biology modeling?

Another example of a software tool was motivated by a collaborative research project among several institutions providing multiple large-scale data sets. In order to make these data easily accessible and searchable for all parties involved, a web-based interface for the data was developed. Upcoming questions during development were:

Which functionality enables the initial assessment and analysis of big data sets? How can data are made available to users from multiple institutions independently from their local infrastructure and technical background?

Identification of limiting factors during influenza virus infection

Since the *Spanish flu* in 1918–19 killed 20 to 40 million people worldwide [151], influenza A virus (IAV) infection is known to be a continuous threat to humanity. While the human transportation networks in 1918 were small and the means of transportation extremely slow, todays air transportation system connects every small province airport to the rest of the world in less than six stops [4]. The impact on the spread of infections is tremendous and can be modeled as circular waves on a specific distance measure [20]. Therefore, despite the advances in vaccination, prevention and treatment of IAV infection, the risk of a new pandemic remains high. In order to identify new drug targets and develop universal vaccines, a fundamental understanding of how influenza viruses replicate is critical. In the scope of this thesis, I focus on the first steps of virus infection, which are the entry of IAV into the host cell and the transport of the viral genome into the nucleus. While the mechanisms of virus-endosome fusion have been

investigated experimentally, the impact of altered pH-sensitivity to the success rate of infection remained largely unclear. Using a systems biology approach, I want to tackle the following set of questions:

How can intracellular processes regulate host specificity and strain pathogenicity? Which nodes in the reaction network are limiting for the infection? How do changes in pH-sensitivity of the viral hemagglutinin affect the infectivity of IAV?

Optimizing the treatment of cancer and associated health conditions

The lack of red blood cells is termed *anemia* and represents a major issue in cancer patients, both as a side-effect of chemotherapy or induced by the malignancy itself [174]. Also in patients suffering from chronic kidney disease (CKD), anemia is highly prevalent. Almost every second pre-dialysis CKD patient shows hemoglobin levels below $12\,\mathrm{g\,dl^{-1}}$ [129]. The *standard of care* for anemic patients consist of a treatment either by blood transfusion or by administering an *erythropoiesis stimulating agent* (ESA). While blood transfusions can be very costly and are associated with an increased risk of infections [174], the treatment using ESAs can be realized in self-administration using a pen injection similar to an insulin pen for patients with diabetes [171]. However, the treatment of cancer patients suffering from anemia using ESAs has been discussed controversially as the ESA might stimulate survival and proliferation of cancer cells [78]. Furthermore, even within the dosing guidelines provided by the drug authorities, the dose might be inadequate for patients with an altered responsiveness to the drug and might cause severe damage in risk groups. Here, a well-studied model of receptor trafficking and signaling in response to ESA treatment is utilized to characterize the *in vivo* properties of different ESAs and to infer the number of ESA binding sites for different cell lines. By extending the model with a PK/PD part, the effect of the ESA dose can be simulated and the dosing schedule for individual patients can be optimized on-the-fly. This optimized treatment strategy can potentially contribute to a safe and efficacious ESA treatment of anemic patients. The corresponding research questions are formulated as:

Is there a better way to treat anemic patients? How can an optimal dosing schedule for individual patients be achieved?

The treatment of cancer itself typically relies on surgery, chemotherapy and radiation therapy. Only recently, a new class of anti-cancer drugs entered the field: immunotherapy drugs. These drugs aim to recruit and activate immune cells to fight malignant tissue and therefore might be used in combination with conventional treatments improving the outcome of therapy significantly [109]. The key to success is the identification

of appropriate patient populations for immunotherapy treatment. It was shown that infiltration of the tumor micro-environment by certain immune cell types can serve as a prognostic factor for overall survival [147]. However, the composition of the tumor micro-environment is largely unknown and can only be broken down by histology of tissue samples. In the past decade, high-throughput sequencing technologies evolved rapidly and acquisition of RNA-seq data of tumor biopsies has become a common step in clinical protocols [184]. Bulk gene expression data can be used for obtaining the cellular composition via mathematical deconvolution. Previous deconvolution approaches were based on micro array data and their accuracy was benchmarked on liquid samples such as *in vitro* cell mixtures or whole blood samples [1, 157, 139]. In this project, I investigated how RNA-seq data can be used for deconvolution of solid tumor samples. By adapting existing methods to the new technology, absolute proportions of immune cells and tumor-associated cells can be derived based on reference gene expression profiles from single-cell RNA-seq data. The questions asked here are:

Does the micro-environment of immune cells affect gene expression? How reliably can the immune cell content in tumor tissue be predicted from blood-derived reference profiles?

Thesis structure

In the following, I want to shortly outline the structure of this thesis. In Chapter 2, both the experimental and the computational methods utilized in the scope of this thesis are introduced. Here, the focus is on the basic concepts of the methods rather than on an exhaustive description. In Chapter 3, two customized software solutions for computational biology projects are presented. While the MATLAB based D2D software, presented in Section 3.1, provides efficient implementations of established methods mainly for ODE modeling, the data interface presented in Section 3.2 aims to make big data sets easily accessible for users without a computational background. Next, in Chapter 4, the infection of an organism with the influenza A virus is investigated. By using a combination of ODE and spatial-stochastic modeling, limiting factors of virus entry into the host cell could be determined. Chapter 5 spans two projects that deal with cancer-related modeling approaches to optimize treatment strategies. The first project deals with an optimized ESA treatment in cancer-associated anemia while the second uses computational deconvolution of tumor tissue from RNA-seq data to characterize immune cell infiltration. In the last chapter (Chapter 6), the individual results are brought into a broader context and the projects are embedded into the current developments in the field of systems biology.

2 Methods

2.1 Experimental methods

In this section I want to introduce the biological measurement techniques that have been used to generate the data for the different projects. The aim is to give the reader a very short introduction into the techniques and to provide references to literature where more detailed descriptions can be found.

2.1.1 Fluorescence microscopy

Fluorescence microscopy is used to monitor specific features within a living or fixated cell. Therefore, parts of the cell or individual molecules are labeled directly with a fluorescent molecule (e.g. nucleus staining using DAPI) or labeled indirectly through a conjugated antibody and a fluorescent secondary antibody. Another option is the expression of fluorescent proteins (e.g. *green fluorescent protein*; GFP) coupled to a protein of interest. The sample is then excited at a fluorophore-specific wavelength and the resulting emission at higher wavelength is captured using a camera on the microscope.

A review on the topic providing physical details and explaining recent developments is given in Lichtman & Conchello [118].

2.1.2 Flow cytometry and cell sorting

Flow cytometry – in combination with *cell sorting* also known as *Fluorescence Activated Cell Sorting* (FACS) – is a technique to characterize and sort cell populations according to fluorescent cell markers, for instance antibodies. By applying multiple steps of cell sorting, even cells with a low abundance in the overall population can be separated (e.g. regulatory T cells in whole blood samples).

The method was developed by Fulwyler [62]. The principle is as follows: cells in suspension are entrained into a stream of liquid. The stream is broken into droplets containing

only one cell each. Just before splitting the stream into droplets, fluorescently labeled antibodies are excited by a corresponding laser and, depending on the fluorescence intensity, the droplets are charged accordingly. Cell droplets are then sorted based on their charge through deflection in an electrical field. The process not only sorts the cells but also measures fluorescence and scattering of the light source in two directions to the stream (i.e. *forward scatter* (FSC) and *side scatter* (SSC)). These parameters provide additional information on cell size (proportional to FSC) and surface properties (granularity; proportional to SSC) of the analyzed cells.

While first generation FACS devices could measure three parameters (one fluorescence color and two scattered-light signals), modern devices offer up to 19 parameters (17 fluorescence colors and two scattered-light signals) [153]. Therefore, modern flow cytometry can, for instance, characterize many subpopulations of immune cells in only one experiment. The historical development is reviewed in Herzenberg *et al.* [80]. The more recent development of multi-color devices is described in Perfetto *et al.* [153].

2.1.3 Quantitative immunoblotting

Immunoblotting, also known as *Western blotting*, is a method to detect and quantify specific protein content in a biological sample using antibodies that react with the target protein. The technique was developed by Towbin *et al.* [193] and named in reference to the *Northern* and *Southern blotting* technique to detect RNA and DNA, respectively.

For quantification of relative protein concentrations, cells are lysed and the proteins are sorted by size via gel electrophoresis. Proteins move towards the anode at a speed according to their molecular mass, usually measured in kDa. The resulting bands in the gel can be transferred to a nitrocellulose membrane and the membrane is treated with an antibody against the proteins of interest (*primary antibody*). The concentrations can be detected through an enzymatic process of a reporter linked to the *secondary antibody*. The chemiluminescent or fluorescent signal is then recorded by a CCD camera and quantified from the band intensity of the image. A more extensive description of the method including protocols for the experiment can be found in Yang & Mahmood [215].

As investigated by Schilling *et al.* [177], gel effects can be minimized using a randomized order of lanes. The errors of Western blotting data were shown to be mostly log-normally distributed [104].

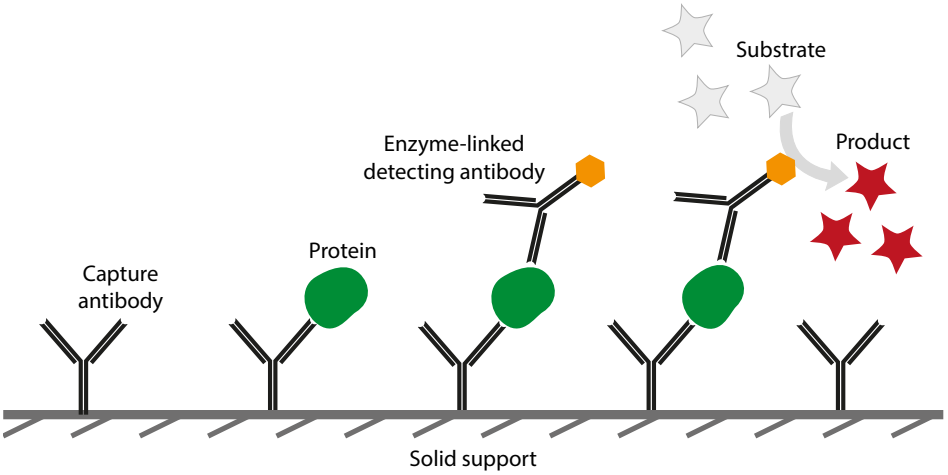

Fig. 2.1.: Schematic representation of a sandwich ELISA. A known quantity of capture antibody is bound to a surface. The sample is added and the protein of interest binds the capture antibody. A secondary detecting antibody is added and binds to the protein. This antibody is linked to an enzyme, which converts an added substrate to a product with specific absorption at a defined wavelength. The absorbance can be measured and quantifies the abundance of the protein.

2.1.4 Enzyme-linked immunosorbent assay (ELISA)

ELISA is an analytical biochemistry assay originally developed by Weemen & Schuurs [209] and Engvall & Perlmann [56]. Using at least one enzyme-linked antibody, the presence of a target protein is made visible through a color change caused by an enzymatic reaction. To asses the concentration quantitatively, absorbance of the converted substrate is measured. In Figure 2.1 the biochemical process of a sandwich ELISA is schematically depicted.

2.1.5 Quantitative real-time PCR (qPCR)

PCR

Using the *polymerase chain reaction* (PCR) technique, single copies of pieces of DNA can be amplified over several orders of magnitude. The process is depicted in Figure 2.2. In a first step, original DNA is denaturated by a temperature increase to $94\,°\mathrm{C}$ to $96\,°\mathrm{C}$. At given annealing temperature, two primers bind to the 3' and 5' end of the target sequence thereby flanking it. A heat resistant polymerase detects the double-stranded regions and starts to elongate the primers according to the complementary strand at an enzyme-specific higher elongation temperature. This temperature cycle is repeated 20 to 40 times until the reaction saturates due to exhaustion of reagents and enzyme.

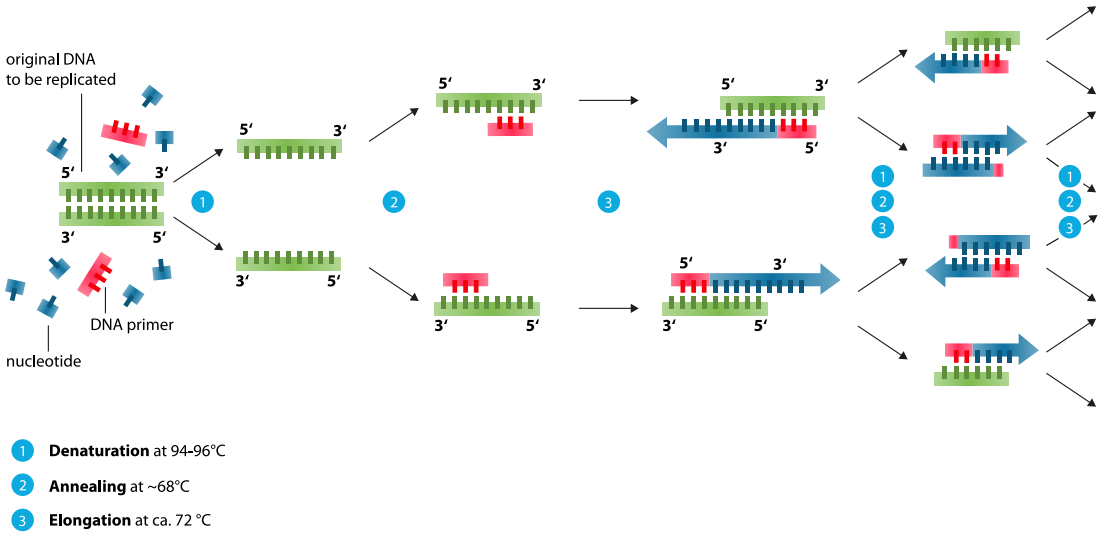

Fig. 2.2.: Scheme of the processes involved during PCR. (1) The original DNA to be replicated (green) is denaturated by a temperature step to 94 °C to 96 °C. (2) Primers (red) bind to DNA at ∼68 °C. (3) Elongation of new double-stranded DNA (blue) at ∼72 °C. Repetition of steps (1)-(3) until saturation is reached. The figure is taken from Enzoklop [57].

.

quantitative real-time PCR

In order to quantify the amplification, the original DNA is labeled using a fluorescent dye. Therefore, in each cycle, fluorescence increases proportionally to the amount of amplified DNA. The cycle number at which the fluorescence raises above a noise threshold corresponds to initial number of copies in the sample.

2.1.6 High-throughput RNA sequencing (RNA-seq)

High-throughput RNA sequencing aims to quantify gene expression of the whole transcriptome. It is based on *next-generation sequencing* of cDNA obtained by reverse transcribing RNA.

Next-generation sequencing (NGS)

High-throughput or *next-generation sequencing* is a technique to sequence large amounts of DNA in parallel. There are multiple technologies available based on different sequencing approaches. Here, I will focus on the *sequencing by synthesis* approach that is used by Illumina/Solexa platforms.

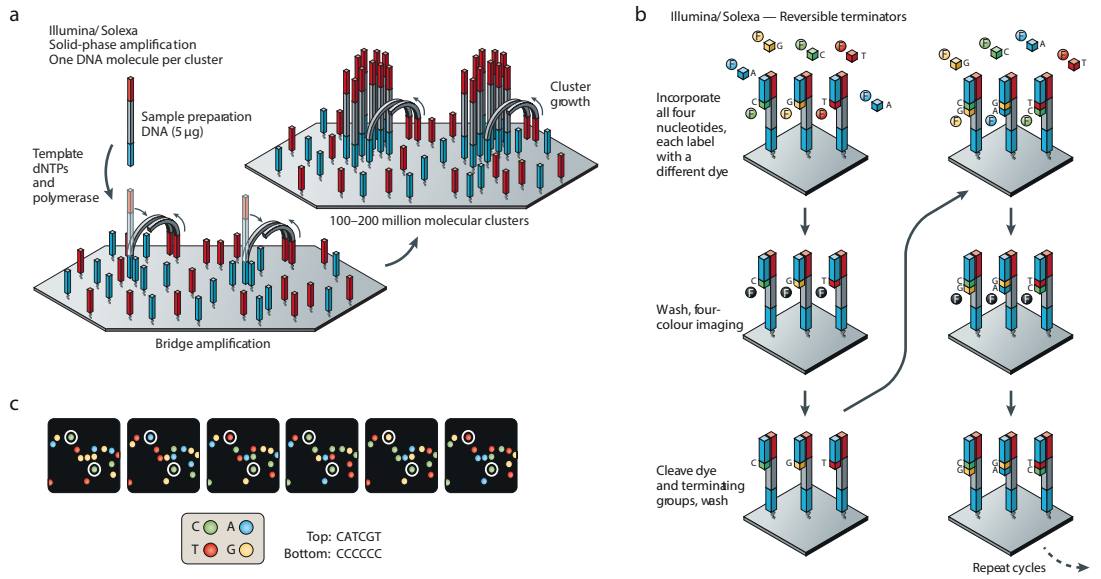

Fig. 2.3.: Next-generation sequencing using the Illumina/Solexa technology. **(a)** Bridge amplification of molecular clusters. **(b)** Sequencing by cycles of binding of labeled nucleotides, washing and four-color-imaging and cleaving of dyes and terminating groups. **(c)** Exemplary imaging pictures for six sequencing cycles. The figure was adapted from Metzker [132].

Firstly, a template library needs to be prepared by breaking the DNA into short fragments (typically $100\,\mathrm{bp}$ to $250\,\mathrm{bp}$) and ligating an adapter set on both ends of the fragment. These single-stranded DNA templates are then hybridized to the corresponding adapters that are covalently bound to a glass slide in a flow cell as shown in Figure 2.3a. Subsequently, the templates are clonally amplified through a *bridge amplification* so that spatially separated template clusters are produced.

After amplification of the templates, the sequencing process is initiated by adding primers that are complementary to the adapter regions initiating the binding of polymerases to double-stranded regions as shown in Figure 2.3b. In each cycle, a set containing of all four labeled nucleotides is added and incorporated into the new DNA strand. These nucleotides are 3'-blocked so that each cluster can only incorporate one base. After washing all unincorporated bases, four-color imaging of the slide is performed, fluorophore and blocking group are cleaved and the next cycle is started by adding new nucleotides. This cycling is repeated until the whole fragment is imaged or a predefined read length is reached. An exemplary imaging sequence of DNA template clusters is depicted in Figure 2.3c. Each cluster emits one color depending on the lastly incorporated labeled nucleotide. More in-depth descriptions of the commonly used technologies are available in reviews by Metzker [132], Goodwin *et al.* [71], and Mardis [125].

RNA-seq

As stated above, for RNA-seq RNA is converted to cDNA by reverse transcription before NGS can be applied. Therefore, only a few additional details need to be introduced here. As a first step in data generation, mRNA or total RNA is fragmented and is reverse transcribed to a cDNA library. To ensure that also less abundant transcripts are detected, ribosomal RNA and very abundant transcripts are removed beforehand. Also, poly(A) selection is effective to remove non-coding mRNA from eukaryotic cells. Unlike in the case of genome sequencing, assembling the whole transcriptome from short reads is not trivial because transcripts can share some exons making mapping from reads to transcripts ambiguous. Therefore, longer reads are preferred but the most common NGS technologies (e.g. Illumina) only provide reads of $75\,\mathrm{bp}$ to $150\,\mathrm{bp}$. This problem can be alleviated by using a paired-end protocol, where overlapping reads from both ends of the cDNA are merged computationally [126].

single-cell RNA-seq

For single-cell transcriptomics, a variant of the RNA-seq method can be applied: Solid tissue samples are dissected into small cubes and digested to obtain a single-cell suspension. RNA from individual cells is extracted and reverse-transcribed to cDNA. To be able to determine the cell of origin after sequencing, an $8\,\mathrm{bp}$-barcode is added to the 5'-end of the template. As the initial amount of RNA and therefore of cDNA is very small, amplification of cDNA by PCR is needed. However, it was shown that amplification can cause a severe bias as not all transcripts are amplified equally. To overcome this bias, the use of a second tag, called *unique molecule identifier* (UMI), typically $5\,\mathrm{bp}$ long, was proposed [99] and was successfully used to correct for the amplification bias [91]. With these two tags, the cell-specific barcode and the molecule specific UMI, an absolute molecule count per cell can be established [91].

2.2 Data analysis

High-throughput technology can produce enormous amounts of data that is often not directly interpretable. Data from RNA-seq experiments, for instance, provide short reads of cDNA snippets that need to be mapped to a reference genome and require correction for miscellaneous sources of errors. Here, I briefly want to introduce the concepts used to process and analyze these data and provide some useful references to respective literature.

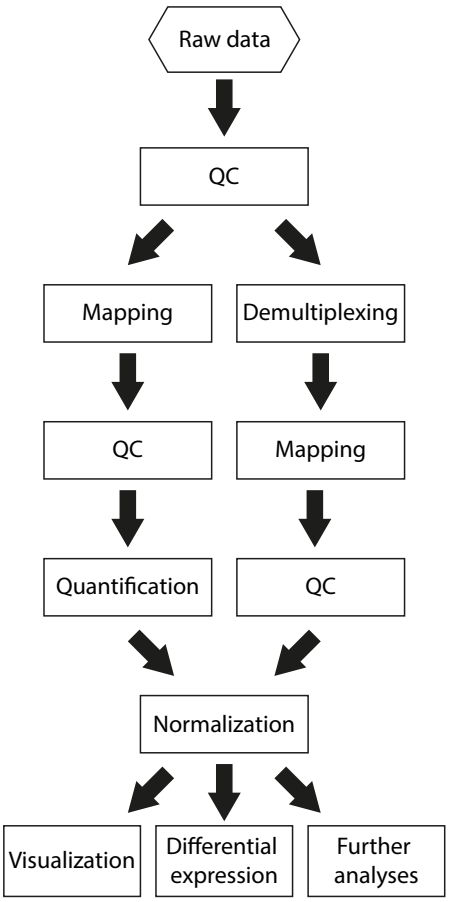

Fig. 2.4.: Flow chart of a typical RNA-seq processing pipeline. The left branch represents the pipeline for bulk data, the right branch for single-cell data.

2.2.1 RNA-seq data processing

The output of an RNA-seq experiment is stored in a text-based *FASTQ* file. For each read, it contains a read identifier (instrument ID, read length, flow cell lane etc.), the nucleotide sequence and the corresponding quality scores. These raw data files need to be processed into a quantity that is proportional to the number of RNA molecules present in the sample. The individual steps of such a processing pipeline are introduced in the following paragraphs.

The typical steps for processing single-cell and bulk RNA-seq data are depicted in a flow chart in Figure 2.4. For each step, various software solutions have been implemented and many institutions developed custom in-house scripts for data processing. I therefore will only mention the basic concept of each step and refer to some prominent solutions.

Quality control (QC)

In the first *quality control* step, the raw data is scanned for low quality reads, for instance reads with biased base composition, that can be filtered out prior to mapping to avoid potential mismatches when mapping reads to the reference genome. A second QC step is performed after mapping and typically includes analysis of GC-content bias, PCR bias, nucleotide composition bias, sequencing depth, strand specificity, coverage uniformity and read distribution over the genome structure. Available tools for QC are provided e.g. by Wang *et al.* [205] and DeLuca *et al.* [42].

Demultiplexing

In the *demultiplexing* step, barcodes identifying individual cells or samples are removed and the data are split into separate files accordingly. If UMIs were used for single-cell RNA-seq experiments, these identifiers are extracted as well and stored as meta data.

Mapping short-reads to the transcriptome

The mapping step serves to align the short reads obtained by sequencing to a reference genome or transcriptome. For large transcriptomes, short reads can match multiple locations in the genome and therefore an unambiguous alignment is not given [207]. Using a paired-end protocol, the read-length can be increased alleviating the problem of multiple matches across the transcriptome.

Prominent alignment algorithms are e.g. STAR [47], Bowtie [112] and Bowtie 2 [111]. Other algorithms like TopHat [194] enable identification of splice junctions from short-read RNA-seq data.

Quantification of read counts

After mapping of short-reads to the reference genome, for each gene or isoform, the occurrence can be counted. These *raw counts*, however, are highly biased by gene length and library size. Typically, these biases are corrected by using measures like *reads per kilobase per million reads* (RPKM) that aim to provide a quantity that is relative to the molar RNA content in the sample. However, it has been shown that the RPKM measure is inconsistent among samples and a corrected measure, i.e. *transcripts per million* (TPM) was suggested by Wagner *et al.* [202]. Modern quantification algorithms like RSEM [115] provide results in this format making data easier to interpret.

Upcoming approaches for integrated mapping and quantification

More recently, novel approaches for combined alignment and quantification of RNA-seq data have been developed [150, 19]. These algorithms are extremely efficient as the actual alignment, which is computationally very expensive, is replaced by a rough *pseudo-alignment* or *lightweight-alignment* that suffices for the integrated quantification strategy. Using these new methods, it is possible to quantify raw short-read data within minutes on a regular laptop computer while former alignment and quantification strategies take hours to days and require big capacities of memory [19]. Interestingly, when comparing quantification results of the novel algorithms with established methods, the overall quality seems to be comparable.

Normalization

While the quantification result measured in TPM is consistent for all samples from the experiment, comparing samples across different experiments and technologies requires additional normalization strategies. Common methods involve *median* or *upper quartile normalization*, assuming that most genes are expressed independently of the experimental condition or origin of tissue. Also revised lists of house-keeping genes can be used for normalization [52]. A comparison of normalization approaches was performed by Li *et al.* [117]. For single-cell sequencing data, normalization methods have been compared by Ding *et al.* [46].

2.2.2 Dimensionality reduction

Data obtained by high-throughput technologies such as RNA-seq are often hard to interpret due to their high dimensionality. Instead of looking at gene expression levels for individual transcripts, clusters of genes with related functionality and differences of these clusters among cell types or patients are of interest. To extract and visualize these properties, dimensionality reduction can be a valuable tool, reducing the space of $\sim 20\,\mathrm{k}$ genes to a low-dimensional map.

Principle Component Analysis (PCA)

Principle component analysis [152] is a method for dimensionality reduction. It uses an orthogonal transformation to transform a set of correlated variables into linearly uncorrelated variables that are called principal components. These components are

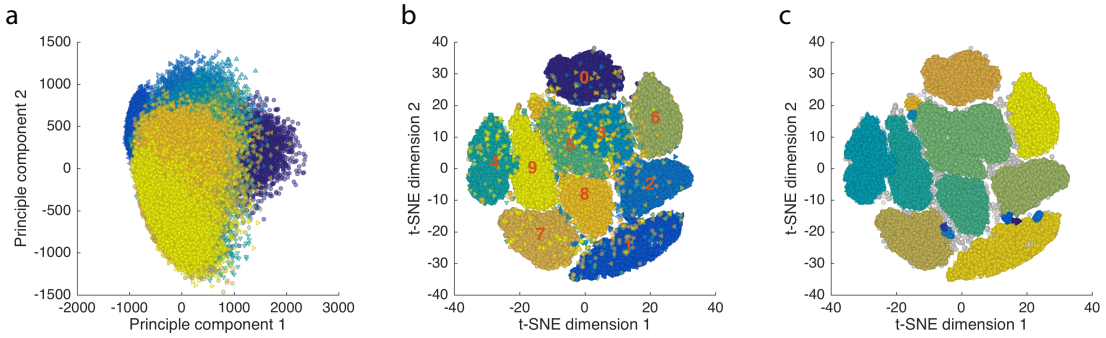

Fig. 2.5.: Example for PCA, t-SNE and DBSCAN clustering applied to the MNIST (*Modified National Institute of Standards and Technology*) database of handwritten digits [114]. **(a)** The first two principle components do not fully separate the points by the digits, as depicted by colors. **(b)** Two-dimensional t-SNE mapping shows distinct clusters for each digit. **(c)** DBSCAN clustering on the unlabeled t-SNE map can correctly identify most of the clusters. Grey dots indicate points without cluster assignment. The parameters for DBSCAN were set to `Eps=1.0` and `MinPts=50`.

sorted by the amount of variance they explain. Therefore, often, only the first few components are analyzed in order to identify interesting features in high-dimensional data sets. In Figure 2.5a, the first two principal components of a typical test data set are shown. In Section 5.3, PCA was applied before decision tree classification and t-SNE mappings.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

A recently developed approach for mapping a high-dimensional data set onto a low-dimensional map is *t-distributed stochastic neighbor embedding* [196]. It maps each high-dimensional feature, such as the gene expression profile of an individual cell, to a two- or three-dimensional point in such a way that similar features result in nearby points and dissimilar features are represented by more distant points. The resulting map can be represented as a scatter plot. In Figure 2.5b, the result of a t-SNE mapping of the MNIST database is depicted. In Section 5.3, t-SNE was used to identify clusters of single cells belonging to the same cell type.

2.2.3 Clustering

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a clustering algorithm originally developed by Ester *et al.* [58]. The algorithm groups neighboring points

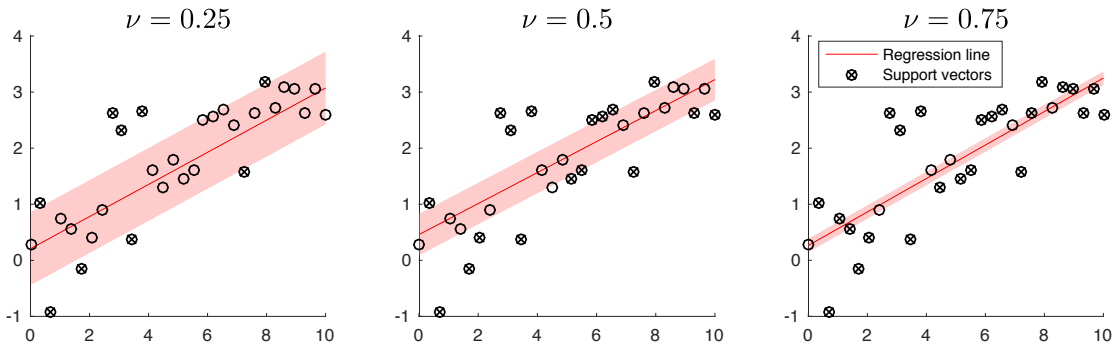

Fig. 2.6.: Example of a ν-SVR on two-dimensional data. Depending on the choice of ν, the number of support-vectors and thereby the width of the ε-insensitive tube is varied, as depicted by the shaded area.

depending on their density and labels outliers without direct neighbors. It depends only on two parameters, `Eps` and `MinPts`, that determine both, the sensitivity to noise and the number of points per cluster. Therefore, for each point, the number of points within an `Eps`-radius are counted. If at least `MinPts` are within distance `Eps`, the point is defined as a core point. All core points that are mutually density-connected, i.e. that are reachable from one another, form a cluster.

The result of a DBSCAN clustering on the t-SNE transformed MINST data set is shown in Figure 2.5c. Most clusters, representing individual digits, can be assigned correctly. In Section 5.3, the DBSCAN clustering algorithm was applied on the t-SNE map of single-cell sequencing data to identify clusters of single cells belonging to the same cell type.

2.2.4 Support Vector Regression (SVR)

Support vector regression is an algorithm based on the supervised machine learning method *support vector machine* (SVM). Other than SVM, which is used for classification by defining separating planes in the high-dimensional space, SVR is used for regression of high-dimensional data. The kernel used for regression can be either linear or non-linear and various strategies for the choice of the parameter ε can be applied. When using an ε-insensitive loss function, ε defines a range around the regression hyper-plane that does not influence the regression result. Depending on the size of ε, the SVR becomes more sensitive or robust to noise.

In Section 5.3, the ν-SVR algorithm Chang & Lin [25] with a linear kernel was used for deconvolution of sequencing data from bulk patient samples. In contrast to ε-SVR, instead of controlling the ε-insensitive range directly, the number of support vectors is regulated through the choice of ν. Therefore, a small value of ν corresponds to a

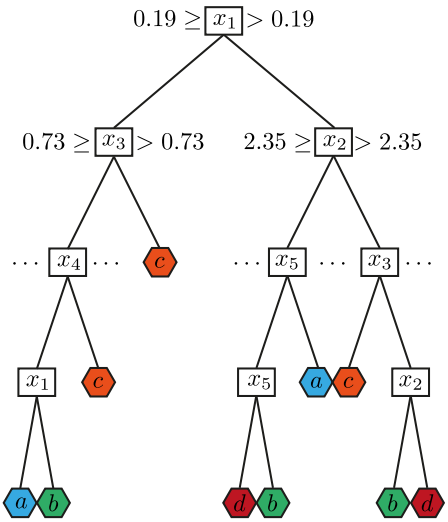

Fig. 2.7.: Scheme of a decision tree classification. A dataset containing five features $x_1, \ldots, x_5$ is classified into four classes a, b, c, d. At each node, a binary decision is made based on the value of one feature. Instead of the features x_i, also the principle components can be used for classification.

wide ε-tube, while a larger ν decreases ε, as shown in Figure 2.6. A detailed tutorial on SVR is given by Smola & Schölkopf [185]. An efficient implementation in C is provided by Chang & Lin [24] in the package `libSVM`, which includes wrappers for Python and MATLAB.

2.2.5 Decision-tree Classification

Classification of unknown data into predefined categories is needed, for instance to assign the cell type of single-cell data based on the gene expression profile. Given a data set with known classes, a decision-tree classifier can be trained. Therefore, in each level of hierarchy, the score of one feature defines a binary decision as shown in Figure 2.7.

In the case of gene expression data, the features correspond to genes. Thus, it can be useful to reduce the high-dimensional feature space by performing a PCA beforehand. The accuracy of the trained classifier can be assessed, for example, by cross-validation, where the training is performed multiple times each time on a subset of the data. Once the classifier is trained, it can be used to classify data with unknown labels. For assignment of cell types for single cells based on RNA-seq data in Section 5.3, a combination of dimensionality reduction, clustering and decision-tree classification was used.

2.3 Dynamic modeling

Many of the methods introduced in this section are described in more detail in

Raue, A., Schilling, M., Bachmann, J., Matteson, A., ***Schelker, M.****, Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U.* & *Timmer, J. „Lessons Learned from Quantitative Dynamical Modeling in Systems Biology“.* PLOS ONE *2013,* ***8****(9): e74335.*

The concepts have been implemented in the Data2Dynamics modeling software that is described in Section 3.1 of this thesis.

2.3.1 Biochemical reaction networks

Biochemical reactions are stochastic events. When the described system is spatially homogeneous (i.e. “well-stirred”) and the number of reacting molecules large, so that many simultaneous reaction events occur, the reactions can be described using *ordinary differential equations* (ODEs) [88].

ODE models

The change over time of the concentration of a substance A can be written as

$$\frac{\mathrm{d}[\mathrm{A}]}{\mathrm{d}t} = v \tag{2.1}$$

where $[\mathrm{A}]$ denotes the concentration of A and v represents the reaction flux. Most general, the form of v can be defined by *mass-action kinetics* where the reaction rate is proportional to the concentration. Hence, for an example reaction

$$\mathrm{A} \rightarrow \mathrm{B}, \tag{2.2}$$

we obtain

$$v = k_1[\mathrm{A}] \tag{2.3}$$

and the temporal change of A is defined by the ODE

$$\frac{\mathrm{d}[\mathrm{A}]}{\mathrm{d}t} = k[\mathrm{A}]. \tag{2.4}$$

k is called the *rate constant* and describes the proportionality of the substance concentration to the reaction flux.

When reactions are catalyzed by an enzyme, *Michaelis-Menten kinetics* [133] can be used, resulting in the ODE

$$\frac{\mathrm{d[S]}}{\mathrm{d}t} = -V_{\max} \cdot \frac{\mathrm{[S]}}{K_{\mathrm{M}} + \mathrm{[S]}} \tag{2.5}$$

where $V_{\max}$ defines the maximum reaction velocity, which depends on the amount of available enzyme and the catalytic activity of the enzyme. K_{M} is the *Michaelis constant* describing the substrate concentration at which the reaction rate is half-maximal. One important prerequisite for using Michaelis-Menten kinetics is that the enzyme concentration is much less than the substrate concentration [180].

Other types of kinetics can be derived from mass action kinetics and can be tailored to the required level of detail. A detailed review on kinetic laws and their applications is given by Tummler *et al.* [195].

In the scope of this thesis, besides mass action, *Hill-kinetics* were utilized:

$$v = \frac{\mathrm{[A]}^n}{(K_{\mathrm{D}})^n + \mathrm{[A]}^n} \tag{2.6}$$

where n is the *Hill-coefficient* defining the steepness of the sigmoidal reaction curve and K_{D} represents the dissociation constant. Hill kinetics describe the cooperative binding of multiple substrate molecules to the enzyme. For $n > 1$, the affinity of a sustrate to bind the enzyme is increased once one molecule is bound [195].

For describing a whole set of ODEs for all compounds involved in the reaction network, we use a more general notation. The *model species* are denoted as $\vec{x} = (x_1, \ldots, x_m)^{\mathrm{T}}$ and the ODEs are defined by

$$\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} = \vec{f_x}(\vec{x}, t) \tag{2.7}$$

where $\vec{f_x}(\vec{x}, t)$ is a function of the reacting species $\vec{x}$ and of the time t.

Stochastic models

When the number of molecules in a reaction compartment is very small, the assumptions for describing the model dynamics using ODEs do no longer hold true. Thus, a stochastic simulation of the model is needed reflecting the stochastic nature of the underlying chemical reactions. Systems that require a stochastic description are, for instance, the dynamics of gene expression on a single-cell level as investigated both experimentally

and theoretically by Elowitz *et al.* [55]. A comprehensive review on the topic is provided by Wilkinson [211].

While for ODE models, the model simulations for a given set of parameters always result in the same trajectories (as the system is *deterministic*), each *realization* will be different for stochastic models. Because the reactions need to be calculated on a single-molecule level, the simulation becomes computationally more expensive for stochastic systems. The *Gillespie algorithm* [68] provides an efficient way to simulate individual realizations of these systems exactly while reducing the computational effort.

Another type of stochastic simulation is required when describing the movement of individual particles, such as virions, in a defined geometry, where spatial aspects need to be considered. Here, the stochastic simulation determines the direction, step-size and time point of the movement rather than the probability of a chemical reaction and the "well-stirred" assumption is no longer needed. On specific geometries, important characteristics like the *mean first passage time* (MFPT) can be calculated analytically [108, 179, 65].

Reaction-diffusion models

Reaction-diffusion models are used when the system under investigation is spatially inhomogeneous, meaning that the diffusion is not sufficiently fast to make the system "well-stirred", and the molecule numbers are too small for a deterministic description of the chemical reactions. Simulation of such models is achieved by discretizing the space into small compartments where the „well-stirred" assumption is valid again. The transition from compartment to compartment can be simulated using a variant of the Gillespie algorithm [212]. Therefore, individual simulations result in different trajectories through the geometry but reactions within a compartment, such as the dimerization of two molecules, follow mass action kinetics.

Modeling frameworks like STEPS [212, 79] and MesoRD [59] provide the means to run simulations on complex, user-defined geometries and offer various features for analysis and visualization. In Section 4.2, STEPS was utilized to describe the diffusion of individual viral genome complexes through the cytosol of the host cell.

2.3.2 Parameter estimation

ODE models of biological systems aim to describe dynamics of the model species over time. As introduced in Section 2.3.1, the differential equations describing biochemical

reactions comprise parameters such as rate constants and Michaelis constants. The values of these parameters may depend on experimental conditions and values from literature are therefore only of limited use.

Experimental data quantifying, for instance the cellular response to an external stimulus, can be utilized to infer unknown model parameters. The parameters to be inferred might also comprise scaling parameters that account for an indirect experimental observation of the quantity of interest or parameters of an error model, as introduced in the following paragraphs.

Observables

Experimental measurements of biological species, such as proteins, often provide only relative abundances, as in case of Western blotting or ELISA (see Section 2.1.3 and Section 2.1.4). In order to link data to the model, so-called *observable functions* are required that define the projections of the model species which are observable experimentally. Typical examples are sums of proteins in different phosphorylation states combined with a scaling parameter accounting for the relative nature of the measurement technique.

More generally, the model species $\vec{x}$ are mapped to an experimental observation through a function $\vec{f_y}$:

$$\vec{y}(t, \theta) = \vec{f_y}(t, \vec{x}(t, \theta_x), \theta_y), \tag{2.8}$$

where θ_x denotes the parameters occurring in the ODEs $\vec{f_x}$ and θ_y are parameters in the observables $\vec{f_y}$ such as scaling parameters.

Error models

The choice of the correct error model is crucial for reliable parameter estimation results. In biology, multiple replicate measurements are generated to obtain an estimate of the margin of fluctuation. These replicates are then used to calculate the mean and the standard deviation for each time point. This assumes that two criteria are met: (1) the uncertainty of the data is normally distributed and (2) the sample size n is large enough to gain a meaningful estimate of the mean and the standard deviation [164].

These prerequisites, however, are not always met, for instance when only a single time course measurement is available as is the case for some observables in Section 4.2. Using a parametric error model, that is fitted simultaneously with the model parameters, can overcome this issue. Therefore, individual replicates are included and the error model is

defined according to the measurement technique, for instance log-normally distributed errors for Western blot measurements [104].

The general form of the parametric error model reads as

$$\vec{\sigma}(t,\theta) = \vec{f_\sigma}(t, \vec{y}(t,\theta), \theta_\sigma,) \tag{2.9}$$

where θ_σ denote the parameters of the error model $\vec{f_\sigma}$. Based on a simulation study, it was shown that this approach yields more accurate parameter estimation results as in case of preprocessed data [164]. For simplicity, in the following only the one-dimensional case is shown and vector notation can be waived. The multidimensional case works analogously.

Maximum likelihood estimation

The quality of a model simulation with respect to measured data can be assessed by the *likelihood* [204]. For normally distributed measurement noise, the likelihood reads as

$$\mathcal{L}(y^\dagger|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma(t_i,\theta)}} \exp\left[-\frac{1}{2}\left(\frac{y_i^\dagger - y(t_i,\theta)}{\sigma(t_i,\theta)}\right)^2\right] \tag{2.10}$$

where $y_i^\dagger$ denotes the i-th of n data points, $y(t_i,\theta)$ represents the corresponding model observable at time t_i and $\sigma(t_i,\theta)$ denotes the error model. When including more than one replicate, t_i can have multiple indices for the same time point.

By taking the logarithm of (2.10), it follows

$$-2\mathrm{LL}(\theta) = \underbrace{n \cdot \ln 2\pi}_{\text{const.}} + \sum_{i=1}^{n}\left[2\ln\sigma(t_i,\theta) + \left(\frac{y_i^\dagger - y(t_i,\theta)}{\sigma(t_i,\theta)}\right)^2\right] \tag{2.11}$$

with $\mathrm{LL}(\theta) := \ln\mathcal{L}(y^\dagger|\theta)$ denoting the *log-likelihood*.

Maximum likelihood estimation aims to maximize the likelihood $\mathcal{L}$ by systematically varying the parameters θ. As the logarithm is a monotonous function, this is equivalent to minimizing (2.11). For this purpose, numerical optimization algorithms can be utilized. The concept of numerical optimization is introduced in Section 2.4.2.

When estimating a parametric error model simultaneously with the model parameters, an additional correction term in (2.11) is needed to account for the reduced degrees of freedom

$$\alpha = \frac{n}{n-\ell} \tag{2.12}$$

with n denoting the number of data points and $\ell := \#\theta_{\mathrm{Fit}} - \#\theta_\sigma$ representing the number of fitted parameters minus the number of parameters comprised in the error model.

This term can be derived in analogy to Bessel's correction for the standard deviation of a finite sample, i.e.,

$$\mathrm{SD} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.13}$$

where $\bar{x}$ represents the sample mean [14].

The full log-likelihood then reads as

$$-2\mathrm{LL}(\theta) = \text{const.} + \sum_{i=1}^{n} 2\ln\sigma(t_i,\theta) + \sum_{i=1}^{n}\frac{n}{n-\ell}\left(\frac{y_i^\dagger - y(t_i,\theta)}{\sigma(t_i,\theta)}\right)^2. \tag{2.14}$$

The *maximum likelihood estimator*, which is the parameter set yielding the highest likelihood value, is denoted by $\hat{\theta}$ and is defined as

$$\hat{\theta} := \arg\max_{\theta} \mathrm{LL}(\theta) \tag{2.15}$$

or analogously

$$\hat{\theta} := \arg\min_{\theta}[-2\mathrm{LL}(\theta)]. \tag{2.16}$$

2.3.3 Parameter identifiability

The identifiability of a model parameter describes how well the parameter can be determined based on model structure and available data. Multiple approaches for identifiability analysis have been developed and were compared e.g. by Raue *et al.* [159]. Here, the focus lies on the profile likelihood approach [203, 162] that allows for the identification of *structurally* and *practically* non-identifiable parameters and provides confidence intervals for the estimated parameters.

The *profile likelihood* (PL) is defined as

$$\mathrm{PL}(\theta_i) = \min_{\theta_{j\neq i}}[-2\mathrm{LL}(\theta)] \tag{2.17}$$

where θ_i is the model parameter of interest and $\mathrm{LL}(\theta)$ denotes the log-likelihood as defined in (2.11).

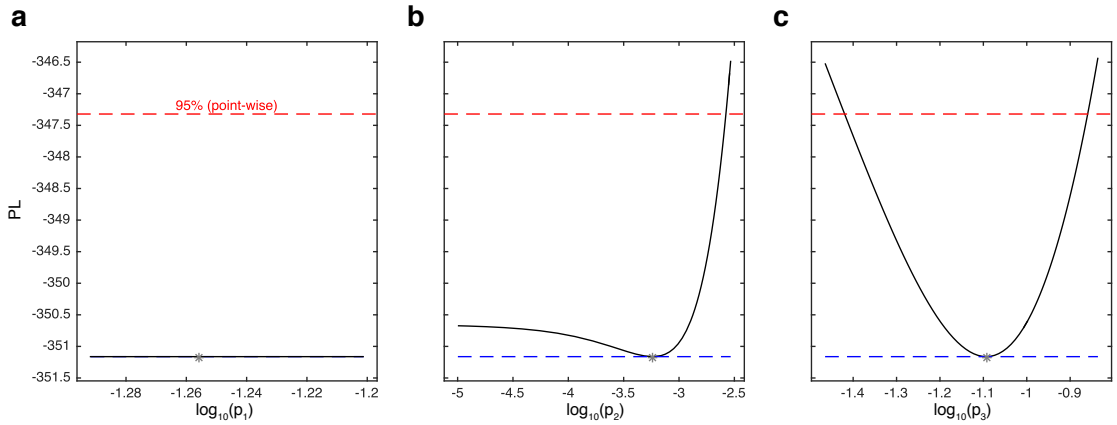

Fig. 2.8.: Identifiability analysis based on the profile likelihood approach. The solid black line indicates the likelihood profile, the gray asterisk shows the value of the best point estimate and the red dashed line indicates the $95\,\%$-confidence threshold (point-wise). **(a)** *Structurally* non-identifiable parameters exhibit a flat likelihood profile. Changes in the value of p_1 do not affect the value of the profile likelihood. **(b)** *Practically* non-identifiable parameters often have half-open confidence intervals due to insufficient quality or amount of data. **(c)** Fully identifiable parameters show parabola-shaped likelihood profiles.

Three exemplary likelihood profiles are shown in Figure 2.8. The shape of the profile indicates the identifiability of the parameter. As shown in panel a), for structural non-identifiable parameters, PL is constant resulting in a flat line. In case of a practical non-identifiability, as depicted in panel b), the data are not sufficient to define both confidence bounds or defines them only on a low level of confidence. The panel c) shows a fully identifiable parameter, where both confidence bounds defined with sufficient confidence.

The confidence intervals can be computed using the *likelihood-ratio test* [201]. Therefore, the reduced model, where the parameter of interest is fixed, is compared with the full model leading to

$$\mathrm{CI}_\alpha(\theta_i) = \{\theta_i | \underbrace{-2\mathrm{LL}(\hat{\theta}) - \mathrm{PL}(\theta_i)}_{\Delta\mathrm{LL}} \leq \mathrm{icdf}(\chi^2_{\mathrm{dof}}, \alpha)\}, \tag{2.18}$$

where α represents the confidence level, icdf denotes the inverse cumulative density function and dof are the degrees of freedom. As the reduced model differs only by one parameter beeing fixed to its point estimate, here $\mathrm{dof} = 1$ and for $\alpha = 0.95$

$$\mathrm{icdf}(\chi^2_{\mathrm{dof}=1}, 0.95) = 3.84 \tag{2.19}$$

and thus

$$\mathrm{CI}_{\alpha=0.95}(\theta_i) = \{\theta_i | \Delta\mathrm{LL} \leq 3.84\} \tag{2.20}$$

or, put another way, as soon as PL is by 3.84 greater than the negative log-likelihood at the optimum, the confidence threshold is crossed.

2.4 Numerical methods

Coupled systems of ODEs, as used to model the dynamics of biological systems, are only in special cases analytically solvable. Fitting unknown rate constants and other parameters to measured data poses a non-linear optimization problem that often exhibits multiple local optima. In this section, the fundamental methods for solving ODEs numerically and for maximizing the likelihood of a model given the data shall be introduced.

2.4.1 Integration of differential equations

Let us start with a simple ODE of the shape

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = f(t, x(t)) \tag{2.21}$$

where $f(t, x(t))$ can be any function of the time t and the variable $x(t)$.

The derivative of $x(t)$ with respect to the time is defined as

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \lim_{h \to 0} \frac{x(t+h) - x(t)}{h}. \tag{2.22}$$

When solving (2.21) numerically, (2.22) is used and the $\lim_{h \to 0}$ is replaced by a small number

$$f(t, x(t)) \approx \frac{x(t+h) - x(t)}{h}. \tag{2.23}$$

Starting from an initial value $x(t = t_0) := x_0$, a recursive formula for (2.23) is obtained

$$x(t_{n+1}) = h \cdot f(t_n, x(t_n)) + x(t_n). \tag{2.24}$$

This numeric approximation of the solution of (2.21) is called the *Euler method* [22]. It represents the simplest approach of an explicit method for solving ODEs numerically. More sophisticated approaches like *Runge-Kutta methods* make use of higher order terms and therefore provide a higher accuracy of the solution. For an extensive description of numerical algorithms for solving ODEs, refer, for example, to Butcher [22].

In this thesis, the numerical solvers from the SUNDIALS suite [81] have been used that are based on the *Adams-Moulton method* and *backwards differentiation formula* (BDF). These implicit methods can also handle *stiff* equations, (i.e. ODEs that involve different time-scales) and offer the simultaneous integration of sensitivity equations (see Section 2.4.2).

2.4.2 Optimization

Estimating rate constants and other parameters comprised in an ODE model based on experimental data is a crucial part of the modeling process. As introduced in Section 2.3.2, the likelihood of a model given the data can be used as an objective function to estimate parameters. This optimization problem needs to be tackled numerically as it can be highly nonlinear with respect to the parameters.

Optimization algorithms

Gradient descent The minimum of an *objective function* $g(z)$ can be found numerically by descending iteratively in the direction of the negative gradient of $g(z)$:

$$z_{n+1} = z_n - \gamma \nabla g(z). \tag{2.25}$$

The step-size γ can be determined in every step of the optimization, for instance using a *line search* [143]

$$\hat{\gamma} = \arg\min_{\gamma} g(z - \gamma \nabla z). \tag{2.26}$$

Newton's method for optimization *Newton's method* can be derived from the Taylor expansion of $g(z)$ around z_n

$$g(z_n + \Delta z) \approx g(z_n) + \nabla g(z_n)\Delta z + \frac{1}{2}\Delta z^{\mathsf{T}} \mathbf{H}_g(z_n)\Delta z \tag{2.27}$$

with $[\mathbf{H}_g(z)]_{ij} = \frac{\partial^2 g(z)}{\partial z_i \, \partial z_j}$ being the Hessian of g.

In order to minimize $g(z)$, a minimum of (2.27) with respect to Δz needs to be found, thus

$$0 = \frac{\mathrm{d}}{\mathrm{d}\,(\Delta z)} \left(g(z_n) + \nabla g(z_n)\Delta z + \frac{1}{2}\Delta z^{\mathsf{T}} \mathbf{H}_g(z_n)\Delta z \right) \tag{2.28}$$

$$= \nabla g(z_n) + \mathbf{H}_g(z_n)\Delta z. \tag{2.29}$$

Hence, a step in direction

$$\Delta z = -[\mathbf{H}_g(z_n)]^{-1}\nabla g(z_n) \tag{2.30}$$

and thus

$$z_{n+1} = z_n - [\mathbf{H}_g(z_n)]^{-1}\nabla g(z_n) \tag{2.31}$$

will decrease the value of g [143].

For so-called *Quasi-Newton methods*, the *Newton step* in (2.30) is scaled by a factor $\gamma \in (0, 1)$ and the Hessian $\mathbf{H}_g(z)$ is replaced by an approximation $\mathbf{B}_g(z)$ [143]

$$\Delta z = -\gamma[\mathbf{B}_g(z_n)]^{-1}\nabla g(z_n). \tag{2.32}$$

Trust-region optimizers A trust-region optimizer approximates g locally by a simple, often quadratic model. The *trust-region* defines the region where the model can be trusted to be an adequate representation of the objective function. In this region, a step minimizing the model is chosen. Therefore, the direction and step-size is chosen simultaneously. If the step is not accepted, the trust-region is shrunk and a new step to minimize the model is chosen [143].

In this thesis, for local optimization mainly the MATLAB function `lsqnonlin` was applied which uses the residuals rather than the value of the objective function. This function is based on a trust-region algorithm [35, 34] and it requires user-supplied gradient information (see also the subsequent paragraph on derivatives). The Hessian is approximated directly by the algorithm.

Local vs. global optimization

In Figure 2.9, the gradient-descent method is shown for a two-dimensional optimization problem. Depending on the initial parameter values, the closest *local optimum* is reached by descending in the direction of the negative gradient. In higher-dimensional systems, such as biochemical reaction networks with several unknown parameters, the problem of local optima can be aggravated. To overcome this issue, either stochastic, *global* optimization algorithms or a multi-start strategy can be applied. While global optimization claims to generally circumvent the problem of "getting stuck" in a local optimum [169], it was shown that multi-start strategies can be more efficient at least for certain classes of models [163]. An overview of different local and global optimization approaches and their application in systems biology is given by [5].

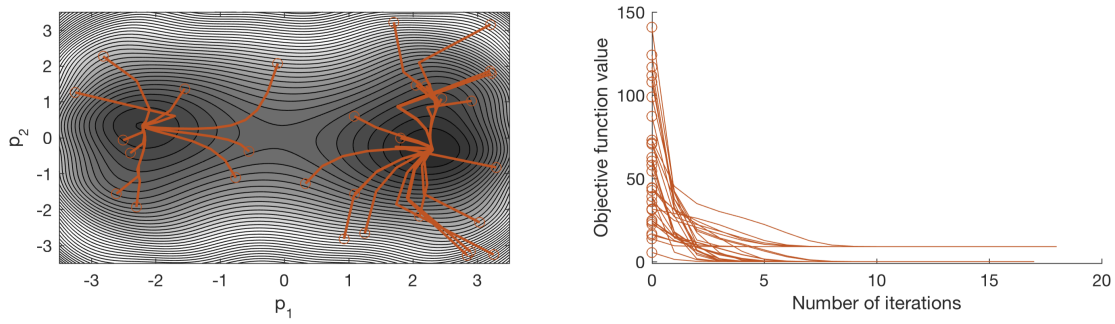

Fig. 2.9.: Local optimization of two parameters p_1 and p_2 from multiple starting points. Lines in orange show the steps of a gradient descent method with a fixed $\gamma = 0.02$. Depending on the initial values of p_1 and p_2, the closest local optimum is reached.

Derivatives

As introduced in the previous section, local optimization algorithms rely on the gradient and Hessian of the objective function to define the direction and step-size of the optimizer. Differentiating the log-likelihood, as defined in (2.14), with respect to θ_j gives

$$\begin{aligned}\frac{\partial \mathrm{LL}(\theta)}{\partial \theta_j} = &-\sum_{i=1}^{n} \frac{1}{\sigma(t_i,\theta)} \cdot \frac{\partial \sigma(t_i,\theta)}{\partial \theta_j} \\ &+ \frac{n}{n-\ell} \sum_{i=1}^{n} \frac{y_i^\dagger - y(t_i,\theta)}{\sigma^2(t_i,\theta)} \left[\frac{\partial y(t_i,\theta)}{\partial \theta_j} + \frac{y_i^\dagger - y(t_i,\theta)}{\sigma(t_i,\theta)} \cdot \frac{\partial \sigma(t_i,\theta)}{\partial \theta_j} \right],\end{aligned} \tag{2.33}$$

where the chain rule and the product rule were applied. Further applying the chain rule leads to

$$\frac{\partial y(t_i,\theta)}{\partial \theta_j} = \frac{\partial y(t_i,\theta)}{\partial x} \frac{\partial x(t_i,\theta)}{\partial \theta_j}. \tag{2.34}$$

Most of the terms in (2.33) can be determined analytically from the model equations. Only the derivative

$$\frac{\partial x(t_i,\theta)}{\partial \theta_j} \tag{2.35}$$

can, in most cases, not be calculated analytically as it depends on the solution of the ODE

$$\frac{\mathrm{d}x(t,\theta)}{\mathrm{d}t} = f_x(t,x,\theta). \tag{2.36}$$

Therefore, (2.35) is typically approximated by *finite differences*, i.e.

$$\frac{\partial x(t,\theta)}{\partial \theta_j} \approx \frac{x(t,\theta+\Delta\theta_j) - x(t,\theta)}{\Delta\theta_j}, \tag{2.37}$$

with $\Delta\theta_j$ being a small perturbation in direction $\vec{e}_j$.

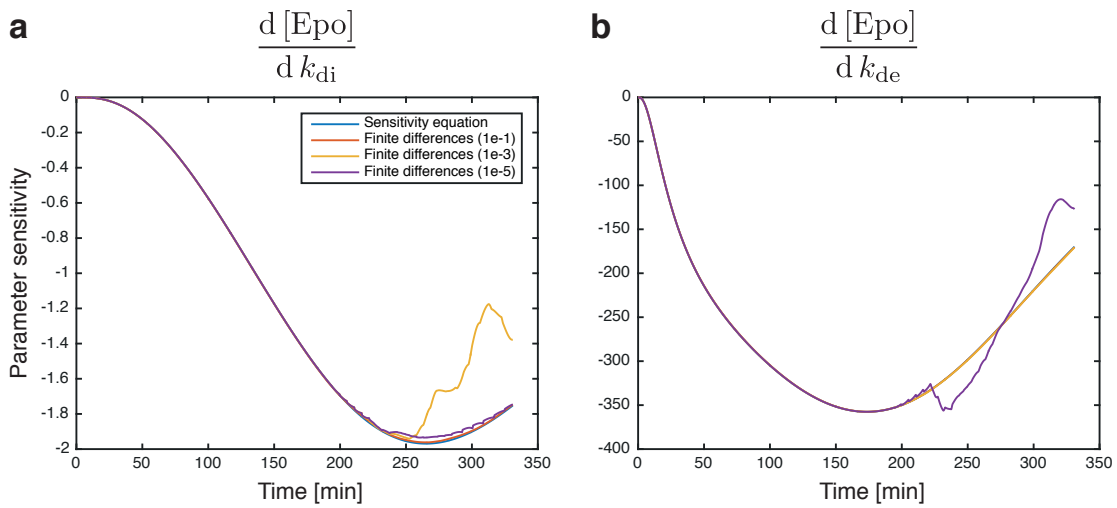

Fig. 2.10.: Comparison of finite differences and sensitivity equation based derivatives. Both approaches are compared for the variable Epo and its derivatives with respect to the parameters **(a)** k_{di} and **(b)** k_{de} from the ESA-EpoR receptor model (Section 5.2). The blue line depicts the solution of the sensitivity equation, red, yellow and purple represent different values of $\Delta\theta$ for the finite differences approach.

When solving the ODE numerically for two parameter sets θ and $\theta + \Delta\theta_j$, the numerical algorithm for solving the ODE takes different time steps for both runs. Subtracting the two solutions can therefore induce numerical noise as shown in Figure 2.10.

Sensitivity equations [44] provide an alternative way of calculating derivatives numerically while avoiding the induction of numerical noise. They are obtained by solving an additional system of ODEs where the derivative in (2.35) represents the new integration variable:

$$\mathrm{SE}_j(t,\theta) := \frac{\partial x(t,\theta)}{\partial \theta_j}. \tag{2.38}$$

Differentiating both sides of (2.38) with respect to the time

$$\frac{\mathrm{dSE}_j(t,\theta)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}x(t,\theta)}{\mathrm{d}\theta_j} \tag{2.39}$$

and switching the differentiation variables leads to

$$\frac{\mathrm{dSE}_j(t,\theta)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}\theta_j}\frac{\mathrm{d}x(t,\theta)}{\mathrm{d}t}. \tag{2.40}$$

With (2.21) and (2.38), it follows the *sensitivity equation*

$$\frac{\mathrm{dSE}_j(t,\theta)}{\mathrm{d}t} = \frac{\partial f_x(t,x,\theta)}{\partial x}\mathrm{SE}_j(t,\theta) + \frac{\partial f_x(t,x,\theta)}{\partial \theta_j}. \tag{2.41}$$

Again, the partial derivatives of f_x with respect to x and θ_j can be calculated analytically. The numeric solution of (2.41), as shown in Figure 2.10, provides a smooth and accurate approximation of the true derivative. An implementation of this method is used in the Data2Dynamics software introduced in Section 3.1.

3 Developing software tools tailored to systems biology applications

In this chapter, I want to highlight the importance of the development of specialized software for systems biology and present two examples of such tools: First, the *Data2Dynamics* (D2D) software is introduced. It provides efficient implementations of well-established methods for data-based dynamic modeling. It enables researchers to include data obtained under various experimental conditions, to estimate measurement uncertainties based on parametric error models and to analyze parameter identifiability. In contrast to other freely available software tools for computational modeling, D2D is constantly evolving through community-driven development on GitHub[1]. The second project consists of a web-based interface for browsing, analyzing and filtering multiple high-throughput data sets acquired within the research project *ViroSign*. It provides fast and platform independent means for gathering from multiple sources and exporting filtered subsets for further analyses.

3.1 Data2Dynamics – A framework for data-based modeling

The Data2Dynamics software is developed in a community effort. The corresponding publication is:

Raue, A., Steiert, B., ***Schelker, M.****, Kreutz, C., Maiwald, T., Hass, H., Vanlier, J., Tönsing, C., Adlung, L., Engesser, R., Mader, W., Heinemann, T., Hasenauer, J., Schilling, M., Höfer, T., Klipp, E., Theis, F., Klingmüller, U., Schöberl, B. & Timmer, J. „Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems.“* Bioinformatics *Nov. 2015,* ***31****(21): 3558–3560.*

[1] `http://www.github.com`

The original parameter estimation benchmark study was published in:

Raue, A., Schilling, M., Bachmann, J., Matteson, A., ***Schelker, M.****, Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U.* & *Timmer, J. „Lessons Learned from Quantitative Dynamical Modeling in Systems Biology“.* PLOS ONE *2013,* ***8****(9): e74335*

My contributions to the project comprise a systematic evaluation of supported SBML features based on the SBML Test Suite [96], improving the SBML import and export capabilities and performing general maintenance and bug fixing. Moreover, I performed the updated benchmark study presented in this section.

Intracellular models of signaling are mostly based on ODEs and therefore require numeric integration routines. The model parameters, describing e.g. initial concentrations or rate constants, are often unknown and existing values from literature are incompatible due to differing experimental conditions. Therefore, efficient and reliable inference of parameters is key for gaining insights from mathematical models of biological systems. As introduced in Section 2.4, a variety of numerical algorithms for solving differential equations and for optimizing the likelihood of a model given experimental data have been developed in the past and implementations in common programming languages are publicly available. Over the last decade, several software tools for model construction, calibration and identification have been developed [178, 83, 219, 124] and have been successfully utilized to model biological systems [see e.g. 154, 187, 218, 12].

Model quantities are often not directly detectable and require mapping through model observables (see Section 2.3.2). Furthermore, data can be acquired under multiple experimental conditions, such as different doses of ligand stimuli, and model variants may apply for data generated using mutants with deletions or over-expression of specific genes of interest. These complex setups demand for flexible modeling frameworks that can be adapted to the specific requirements of individual users. While existing software is often available in a compiled form and providing graphical user interfaces (GUIs) to its users, the software tool presented here, is based on a transparent source code written in the scientific computing language MATLAB (The MathWorks Inc, Natick, MA) and the main functionality does not rely on GUIs. As the name of the software, *Data2Dynamics* (D2D), implies, it has a strong focus on data-based dynamic modeling. Unlike most existing tools, D2D is intended for advanced users of whom many also contributed as authors to the development of individual features.

In the following sections, I want to quickly motivate the use of community based development exemplified by the genesis of D2D, give an overview of the key features of D2D and eventually, provide a benchmark of different algorithms for parameter

estimation on two published models with biological data. For the computational methods D2D is based on please refer to Section 2.4.

3.1.1 Community based development

The development of non-commercial, scientific software is mostly performed by scientists from individual work groups. As permanent positions in science are rare and frequent changes in work locations and institutions are common, the long-term development and maintenance of software cannot be guaranteed and many project can quickly become orphans. Thus, an active community of developers and users from multiple institutions enhances the development of scientific software tools and ensures that the correct operation does not depend on a individual person or work group. This is also crucial for reproducing published models and therefore for reproducibility of science in general. In the following, the genesis of the D2D software is briefly introduced and the strong link to its community-driven development is highlighted.

Version control & community work

The development of the D2D software was initiated by Andreas Raue. During the first two years of development, the number of users and contributors increased continuously and the organization of code became more and more complex. Therefore, in March 2013, the source code of D2D was transferred to the hosting service Bitbucket[2] and the development was tracked using the *version control system* (VCS) Mercurial. This step simplified and accelerated the development dramatically, as all registered users could now contribute to the source code and make their changes available to the community. Also the process of bug fixing is strongly facilitated by the embedded issue tracking system on the web platform.

With the publication of the manuscript [164], the repository was opened to the public and a documentation was created on the wiki of the repository. In spring 2016, the repository, including all revision history and documentation, was migrated to GitHub and the VCS was converted to `git`. While there are only minor differences between both VCS, the choice of the web-platform is crucial to reach potential users and to motivate interested scientists from the field to contribute to the project.

[2] `http://www.bitbucket.org`

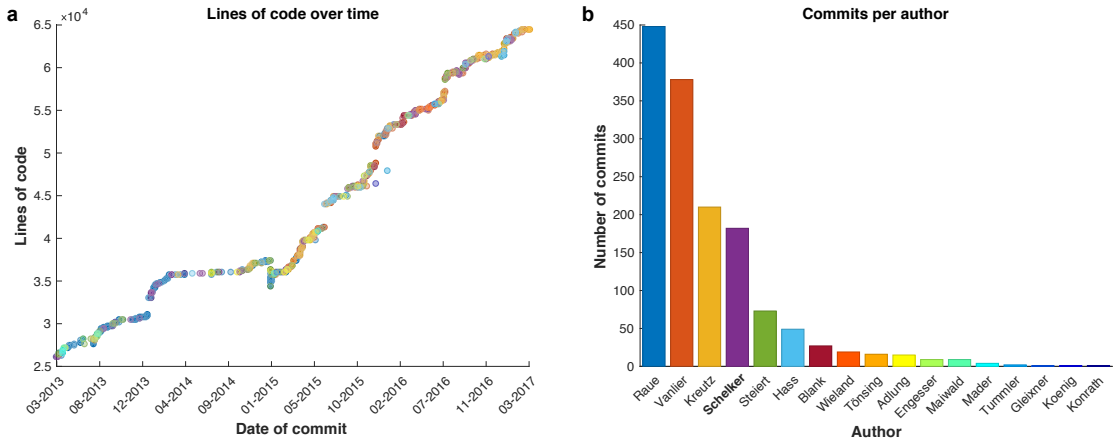

Fig. 3.1.: Evaluation of individual contributions as measured by LOC and number of commits. **(a)** LOC is shown for all parts of D2D written in MATLAB over time. Colored circles indicate the author responsible for the commit. Color-coding as in panel b). Please note that the measure is cumulative and does not represent the LOC added by each commit. **(b)** Number of commits per author. The first five authors account for $\sim 90\,\%$ of the commits.

Statistics and contributions

By using a VCS, the activity of all contributors can be tracked over time. Here, I present a short analysis of the individual contributions over time to emphasize the advantages of a community-based, open development process. The data shown in Figure 3.1 were obtained through a custom shell script making use of the git functions `git log`, `git ls-tree` and `git show`.

In Figure 3.1a, the cumulative measure *lines of code* (LOC) is plotted over time for the whole D2D MATLAB source code. Third party code was excluded from the analysis. The color-coding indicates the author responsible for each commit but does not reflect the extent or importance of individual commits. From the figure it can be seen that from March 2013 to the end of 2014, the development was mainly driven by three authors and the LOC increased only gently from $\sim 25\,\mathrm{k}$ to $\sim 35\,\mathrm{k}$. In 2015 new authors joined and the LOC increased drastically. Therefore, also the functionality of D2D was diversified as authors tend to implement new features required for their own research.

In Figure 3.1b, the overall number of commits is depicted for all contributing developers. Over the last four years, 17 authors made a total of more than 1400 commits to the source code repository. While some authors implemented a few small bug fixes or a specific feature for their projects and therefore had only a few commits, the five most active authors account for $\sim 90\,\%$ and the seven most active developers even for $\sim 95\,\%$ of the commits. This distribution exemplifies that the development is driven by only a few

authors but a larger community contributes sporadically to fix issues and add missing functionality.

Despite the good availability of data from the VCS, it is difficult to find a measure that quantifies the contribution to the functionality, scope and stability of a software, as commits can contain minor or major changes and revision of existing code can even have a negative net LOC despite its utility. Nevertheless, as the software evolves, an increase in LOC, number of commits and contributing authors indicates a steady improvement of existing code and a diversification of features.

3.1.2 Selected features

The core features of D2D, which are simulation of ODE models, parameter estimation and identifiability analysis, have been extended and improved over time. In the following, a small selection of the features of D2D is presented. The emphasis lies on those features that distinguish D2D from existing tools.

SBML import and export

The *systems biology markup language* (SBML) [84] is a model format that enables a detailed description of all species, compartments and reactions. Therefore, SBML import and export functionality is key for the compatibility with other software in the field of systems biology and was included early on.

D2D, as many other tools, makes use of the libSBML library [17] to translate the `xml` format of SBML to a structured array in MATLAB. One difficulty are the different approaches of the model description in SBML vs. D2D: While D2D has a strong focus on signaling models that are comprised of model states, external inputs, time independent parameters and conditions that allow to transform parameters and/or states, in SBML models there are many more possibilities to define the model behavior based on rules, events etc.. Hence, the challenge is to map and translate SBML features to D2D syntax back and forth. However, it should be stated that a full support of all SBML features is not intended as this would go beyond the scope of the software. Even extensive software tools like COPASI [83], that cover a broad field of use cases, provide only partial coverage of the SBML support.

To evaluate the SBML import functionality, the *SBML Test Suite* [96] can be utilized. The current version consists of a total of 1219 models covering the majority of model features that can be defined using SBML. For each SBML model file, a results file indicates

the correct results in terms of the simulated model trajectories and serves as the *gold standard*. Based on the comparison of the gold standard with the D2D simulations, the percentage of supported SBML test models can be compared to other software solutions. For D2D, a total of 273 out of the 1218 test models yielded trajectories with a deviation from the gold standard of less than 1×10^{-6}. Thus, $22.4\,\%$ of the test cases are supported. For established software tools this ratio is substantially higher, as obtained from the *SBML Test Suite Database*[3]: While COPASI [83] supports 1014 out of 1185, libRoadRunner [186] even supports 1072 out of 1185 test cases, corresponding to $85.6\,\%$ and $90.5\,\%$, respectively.

Multi-experiment fitting

In computational modeling, models often need to be consistently calibrated to data obtained under multiple experimental condition such as wild-type vs. knock-out experiments. Thus, the model to be calibrated comprises two conditions that can contain common as well as individual, that is, condition-specific parameters and inputs. When performing parameter estimation, the individual model conditions can be simulated separately but are linked by the common parameters and their contributions to the likelihood.

While this fundamental setting can be difficult to implement in other modeling software, in D2D a separate model variant is automatically generated for each condition defined in the data sheet and no additional configuration is required. As an example for multiple experimental conditions, the different ligands of the Epo receptor in Section 5.2 can be mentioned.

Efficient numerical integration of ODEs

During the process of parameter estimation, the solution of the model simulation for the current set of parameters is compared to the corresponding experimental data. Hence, in each step of the optimization, the ODE system forming the model needs to be solved numerically. Therefore, the numerical integration of the model equations represents a major bottleneck of parameter estimation and needs to be performed as efficient as possible. While scripting languages like MATLAB or Python offer many advantages for fast and human-readable implementation of algorithms, they lack the performance of low-level programming languages such as C or C++. By using an efficient solver

[3] `http://sbml.org/Facilities/Database/Simulator`

implemented in C [81] and connecting it to D2D through MATLAB's `mex` compiler, the numerical integration can be accelerated by up to two orders of magnitude [197]. Also, the `CVODES` solver of the SUNDIALS package [81] provides the option of solving the sensitivity equations simultaneously with the model equations and therefore enables the efficient calculation of derivatives of the likelihood as required by deterministic optimizers (see also Section 2.4.2).

Multi-start deterministic parameter estimation

As introduced in Section 2.4.2 and described in more detail by Raue *et al.* [164], *local optima* during optimization can be overcome either by applying stochastic, *global*, optimization or by performing many runs of deterministic, *local*, optimization strategies with different initial parameter guesses. For these guesses *random sampling* can be utilized to cover the parameter space. However, as the parameter space quickly becomes higher-dimensional, more sophisticated strategies like Latin hypercube sampling (LHS) [146] can be applied. In D2D, both strategies are implemented and can be utilized for multi-start deterministic parameter estimation. Which method is favorable depends on the model size (number of parameters) and the shape of the likelihood landscape (i.e. how many local optima exist etc.). In Section 3.1.3, a benchmark study was performed that evaluates the performance of both strategies on two published models with experimental data.

Identifiability analysis

Depending on model structure and data availability model parameters might not be determined unambiguously by the best fit. This property is called parameter identifiability and is introduced in more detail in Section 2.3.3. To study the identifiability of parameters, methods for identifiability analysis have been proposed [77, 162] and implemented in D2D. The concept of likelihood profiles can be extended to model predictions (Prediction profile likelihood; PPL) [106, 105] which are also available in D2D.

Sensitivity analysis and metabolic control analysis

Knowing which parameter has the most influence on a certain model feature is an important aspect for better understanding models in general and in the quest for identifying drug targets [100]. Sensitivity analysis provides exactly this information for each model variable and observable in a dynamic manner. Metabolic control analysis offers a

framework of *control coefficients* and *response coefficients* [74]. While the control coefficients quantify the influence of reactions on the steady-state concentrations, response coefficients describe the impact of parameters on the steady state. An extension of this framework to non-steady state trajectories was developed by Ingalls & Sauro [89]. Both concepts are included into D2D.

3.1.3 Benchmark of sampling strategies and implemented fitting algorithms

As parameter estimation is one of the key applications D2D is designed for, its efficient and robust functioning is essential. As mentioned before, *local optima* represent a big challenge when estimating model parameters in non-linear ODE models. Therefore, the choice of the optimization strategy – whether *global* or *local* – and the numerical algorithm is critical for successful model fitting. In this section, several optimizers implemented in D2D are evaluated based on a benchmark setup using two existing models with measurement data. For the *local* optimizers, the two different sampling strategies for the initial parameter guesses are tested and compared for both benchmark models.

Benchmark models

For evaluating the performance of different optimization strategies, the two published models from Becker *et al.* [12] and Bachmann *et al.* [8] have been employed including the published data sets. The model by Becker *et al.* is described in more detail in Section 5.2.2 of this thesis. It consists of 6 model species describing the interactions of the hormone erythropoietin (Epo) with its receptor. The model is composed of 8 reactions comprising 7 rate constants. Together with the initial conditions for the model species and the observation parameters it sums up to a total of 16 parameters that have to be calibrated based on experimental observations. Only for the initial Epo concentration, prior information is available which is included by an additional term in the likelihood function.

The model by Bachmann *et al.* describes the downstream signaling of the STAT5/JAK2 pathway upon stimulation with Epo. It describes the dynamics of 25 model species through 36 reactions comprising 21 rate constants. The stimulation of the pathway is triggered by the extracellular Epo level which is modeled as a constant *external input*. The total number of parameters is 115 of which 2 are fixed to a constant due to non-identifiability and for one initial condition the value is set to priorly known value with

Tab. 3.1.: Key figures of the two models used for the benchmark study.

Quantity	Becker *et al.*	Bachmann *et al.*
# of model species	6	25
# of reactions	8	36
# of parameters	16	115
# of external inputs	0	1
# of experimental conditions	1	23
# of data points	85	542
average runtime for one simulation [sec]	0.0043	0.1014

a corresponding penalization term in the likelihood. Another level of complexity is added by the number of experimental conditions: time course data, dose-response experiments and over-expression experiments lead to a total of 542 data points over 23 model conditions.

The two models are very different in size (number of parameters and model species) and complexity (number of experimental conditions, complexity of the kinetic rate laws etc.). Therefore, optimizers are challenged in two very different scenarios that reflect common applications in data-based systems biology modeling projects. The key figures are summarized in Table 3.1. The model equations as well as a schematic representation of the reaction network can be found in the original publications [12, 8].

Comparison of sampling strategies

While random sampling is the straight-forward option to generate a sample of initial parameter guesses for multi-start optimization, the coverage of the high-dimensional parameter space cannot be ensured. As introduced in Section 2.3.2 and in Section 3.1.2, sampling using LHS prevents randomly selected initial parameters from being close to each other. Thus, each optimization run starts in a different region of the high-dimensional parameter space. The effect of the sampling strategy on the performance of multi-start optimization has not been investigated on systems biology models. Here, I want to compare the performance of the two sampling strategies based on the two previously introduced benchmark models.

The result is depicted in Figure 3.2 where the left panel shows 1000 fits for the model by Becker *et al.* and the right panel the results based on the model by Bachmann *et al.* On the y-axis, a measure for the goodness of fit is given by the negative log-likelihood ($-2\mathrm{LL}$, see also Equation 2.14), the x-axis indicates the fit index sorted by the likelihood value. A plateau represents a local optimum [164]. The two sampling strategies are marked by blue (LHS) and red (random sampling) circles.

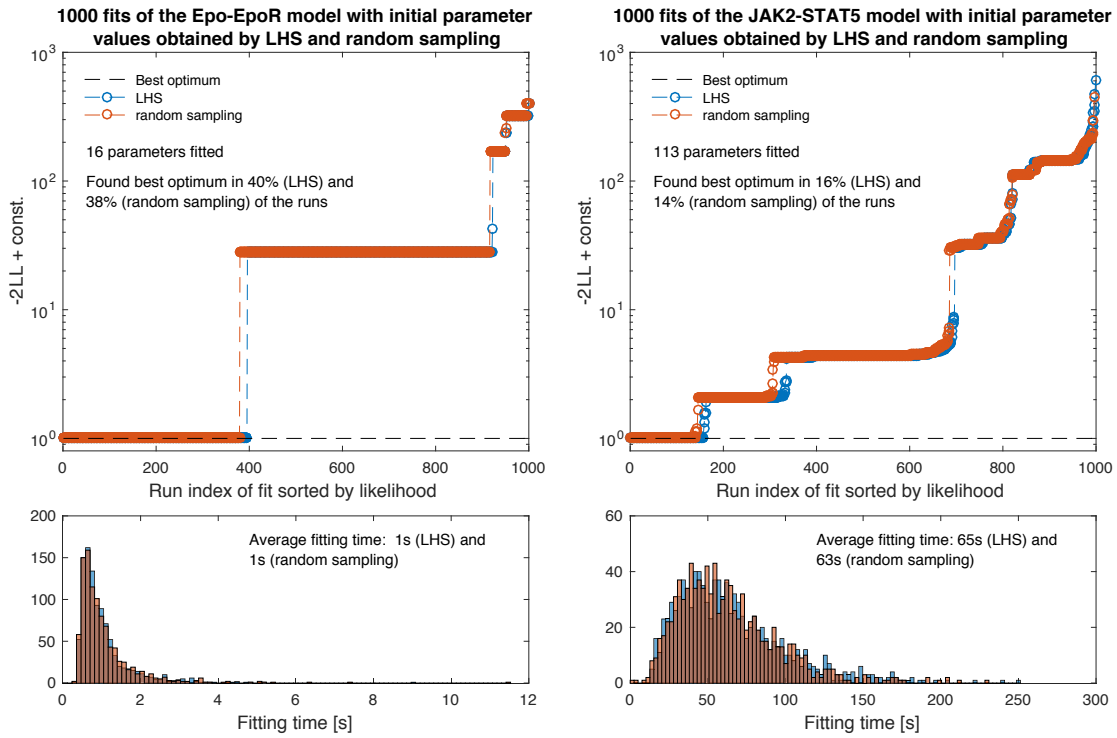

Fig. 3.2.: Comparison of LHS and random sampling based on two published models including experimental data. The upper panels indicate the goodness of fit as measured by the negative log-likelihood on a logarithmic scale vs. the sorted run index. Plateaus in the plot correspond to local minima. Fit based on LHS are depicted as blue circles, fits starting from random sampling are shown in red. The lower panels show a histogram of the fitting times for both sampling strategies and models.

For the model by Becker *et al.* (left panels), only 16 parameters were fitted. Therefore, the parameter space is relatively low-dimensional and the differences in performance, as measured by the number of runs where the best optimum has been found, are small ($40\,\%$ vs. $38\,\%$ for LHS and random sampling, respectively). For the significantly larger model by Bachmann *et al.* (right panels), a total of 115 parameters were fitted. However, despite the high-dimensional parameter space, also here only a small difference in performance can be observed: While the global optimum was found in $16\,\%$ of the runs when using LHS, for random sampling, $14\,\%$ of the runs found the global optimum.

In terms of the average fitting time, both sampling strategies perform equally well ($1\,\mathrm{s}$ for LHS and random sampling for Becker *et al.* and $65\,\mathrm{s}$ vs. $63\,\mathrm{s}$ for Bachmann *et al.*; simulations have been performed on an Intel Xeon X5690 $3.47\,\mathrm{GHz}$ CPU with 12 cores and $94\,\mathrm{GiB}$ RAM).

To summarize, both sampling strategies are able to overcome the problem of local optima when the initial sample size is "adequately" large (D2D uses jackknife resampling [50] to evaluate whether the sample size was large enough). Even though random sampling

Tab. 3.2.: Optimization algorithms used in the benchmark study.

#	Name	Implementation
1	`lsqnonlin`	MATLAB
2	`fmincon`	MATLAB
3	`STRSCNE`	[13]
4	`arNLS`	[161]
5	`fmincon_as_lsq`	MATLAB
6	`arNLS_SR1`	[161]
7	`NL2SOL`	[43]
8	`TRESNEI`	[136]
9	`Ceres`	[2]
10	`lsqnonlin_repeated`	MATLAB
11	`fminsearch`	MATLAB
12	`patternsearch`	MATLAB
13	`patternsearch_hybrid`	MATLAB
14	`particleswarm`	MATLAB
15	`simulannealbnd`	MATLAB
16	`geneticalgorithm`	MATLAB

performs slightly worse, there is one important advantage as compared to LHS when it comes to resampling, i.e., increasing the sample size retroactively: While for LHS the size of the sample cannot be enlarged without starting over, for random sampling the sample can be extended without losing the existing optimization runs.

Benchmark of optimization algorithms

Next, I want to evaluate how different optimizers perform on both benchmark models. The performance of several optimization algorithms implemented in D2D has been evaluated previously [164]. However, since the publication of this comparison, further algorithms have been made available in the D2D framework. Here, a total of 16 optimizers are compared based on the models from Becker *et al.* [12] and Bachmann *et al.* [8]. All optimizers have been used on default settings of their implementation in D2D. The optimizers and the corresponding references for their implementations are summarized in Table 3.2.

For each of the 16 optimization algorithms, 100 fits have been performed with initial parameter guesses obtained by random sampling. The performance of an algorithm was defined as the percentage of fits that converged to the global optimum. Furthermore, the time needed to perform one fit was captured. The results for both models are depicted in Figure 3.3.

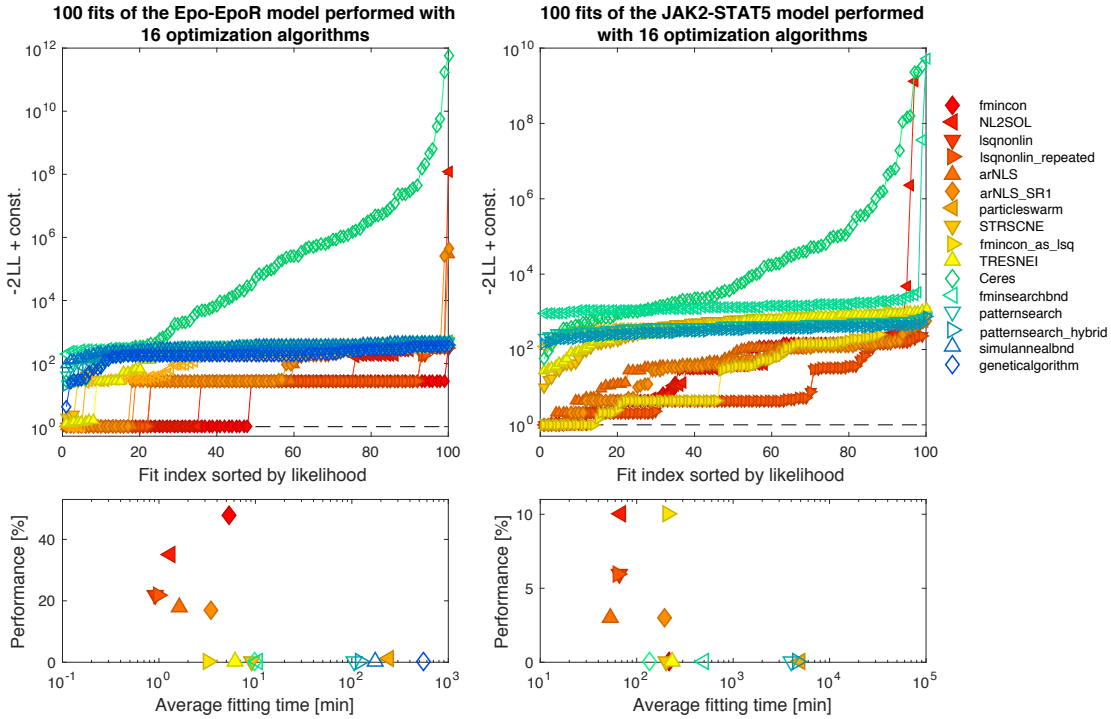

Fig. 3.3.: Comparison of 16 optimization algorithms based on two published models including experimental data. The upper panels show the fitting results for all algorithms as quantified by negative log likelihood. Plateaus correspond to local optima. The best optimum was centered to 1. The lower panels indicate the overall performance of each algorithm vs. the average fitting time. High performance and low time values are desired.

As introduced in the comparison of sampling strategies, in the upper panels, the likelihood is plotted against the sorted fit index so that plateaus, referring to local optima, can be identified. In the lower panels, the performance is plotted against the average fitting time. This diagram allows to evaluate and compare the different algorithms with respect to their reliability in finding the global optimum and to their computation time.

In case of the Becker *et al.* model, two groups could be identified: The deterministic, mostly derivative-based optimizers performed well while the derivative-free and/or stochastic optimizers all showed a poor performance. The top six strategies reached a performance of 15 % to 50 % with an average time of ∼1 min to 10 min per fit (on an Intel Xeon X5690 3.47 GHz CPU with 12 cores and 94 GiB RAM), the remaining algorithms took up to several hundred minutes per fit and succeed only in one out of 100 runs (`particleswarm`) or not at all. For the Bachmann *et al.* model, a similar behavior could be observed. The maximal performance of 10 % was only reached by two deterministic algorithms, namely, `NL2SOL` and `fmincon_as_lsq`. The default optimizer in D2D, `lsqnonlin`, succeeded in 6 % of the fits. All derivative-free deterministic and stochastic optimizers failed to reach the global optimum. Due to the increased complexity

of the model, here, the fitting time ranged from $\sim 50\,\mathrm{min}$ to $200\,\mathrm{min}$ for algorithms with successful fits, and from $\sim 100\,\mathrm{min}$ to $5000\,\mathrm{min}$ for the methods with $0\,\%$ performance.

Taken together, the results of this benchmark study confirm the overall trend observed in the previous analysis [164]: Deterministic, multi-start strategies seem to be more reliable than derivative-free or stochastic optimization strategies. The differences in performance for both benchmark models highlight that there is not one best optimizer but rather a group of well-performing methods and their exact performance depends on the complexity of the given model.

3.2 A data interface for the ViroSign project

This work was done in collaboration with Max Flöttmann in the scope of the ViroSign project. The data were provided by the project partners and were taken from existing publications.

The main goal of this project was to make all data sets from the ViroSign project partners, as well as related published data sets, accessible for participating researchers. Based on our discussions with the project partners, we defined the following list of requirements for building a web interface for the data:

1. The data interface needs to be closed for the public yet accessible for all authorized users,
2. existing data sets from publications can be added,
3. the user can combine (`merge`) individual data sets based on the `gene_id`,
4. the column `gene_symbols` can be searched by exact match and by regular expression patterns,
5. the two columns can be plotted as a scatter plot and the searched genes are highlighted in the plot,
6. the table can be sorted by the values of each column (i.e. `fold_change` or `log10_abundance`), and
7. the filtered table can be exported to a `.csv` or `.xls` file.

As a standard tool for data analysis, the statistics software *R* [158] is equipped with most of the requested features. However, the user would need to learn the programming language and have a copy of data and software installed on his local computer in order to access the functionality. Therefore, we made of use the R-package `shiny` [26] which provides an easy-to-use web interface for the user but offers the full functionality of *R* in the background.

We included a total of eight data sets comprising of two screens of *small-interfering RNA* (siRNA) [95, 172], time-resolved SILAC-based proteomics data of IAV infected cells [172], cap-snatching data [183], absolute protein abundance data of multiple cell lines [66] and unpublished protein-protein-interaction screens for the viral proteins NP, NS1 and PB2.

Figure 3.4 shows a screen shot of the web interface. In the collapsible panels on the left, data sets can be selected and filters based on gene names can be applied. Also, a quick overview is provided by plotting two columns against each other as a scatter plot. Further features are different data merging options and selection of columns to be shown in the data table on the right. A download button below the table provides the capability to save the filtered data to a `.xls` or `.csv` file.

3.3 Conclusions

Based on two projects, I showed that software development plays a crucial role for scientific progress in computational systems biology. While the data interface for the collaborative influenza research project ViroSign has a limited functionality and is closed for the public, it provides unpublished data to all project partners and allows them to obtain a first insight into the highly complex data sets. By contrast, the modeling software *Data2Dynamics* is open to the public and shows a growing community of users that also act as developers for novel features.

Specialized software for computational systems biology is mainly developed by researchers at academic institutions. Thus, the long-term persistence and maintenance of the software often relies on individual scientists employed on non-permanent positions. By transferring the source code to open platforms like GitHub, the functionality of the software can be extended by the users themselves and possible issues are often quickly resolved by the community. Nevertheless, the administration of the project is critical to ensure a well-organized progression. Therefore, software development within systems biology research groups should be further professionalized by hiring programmers and computer scientists for the core development and maintenance.

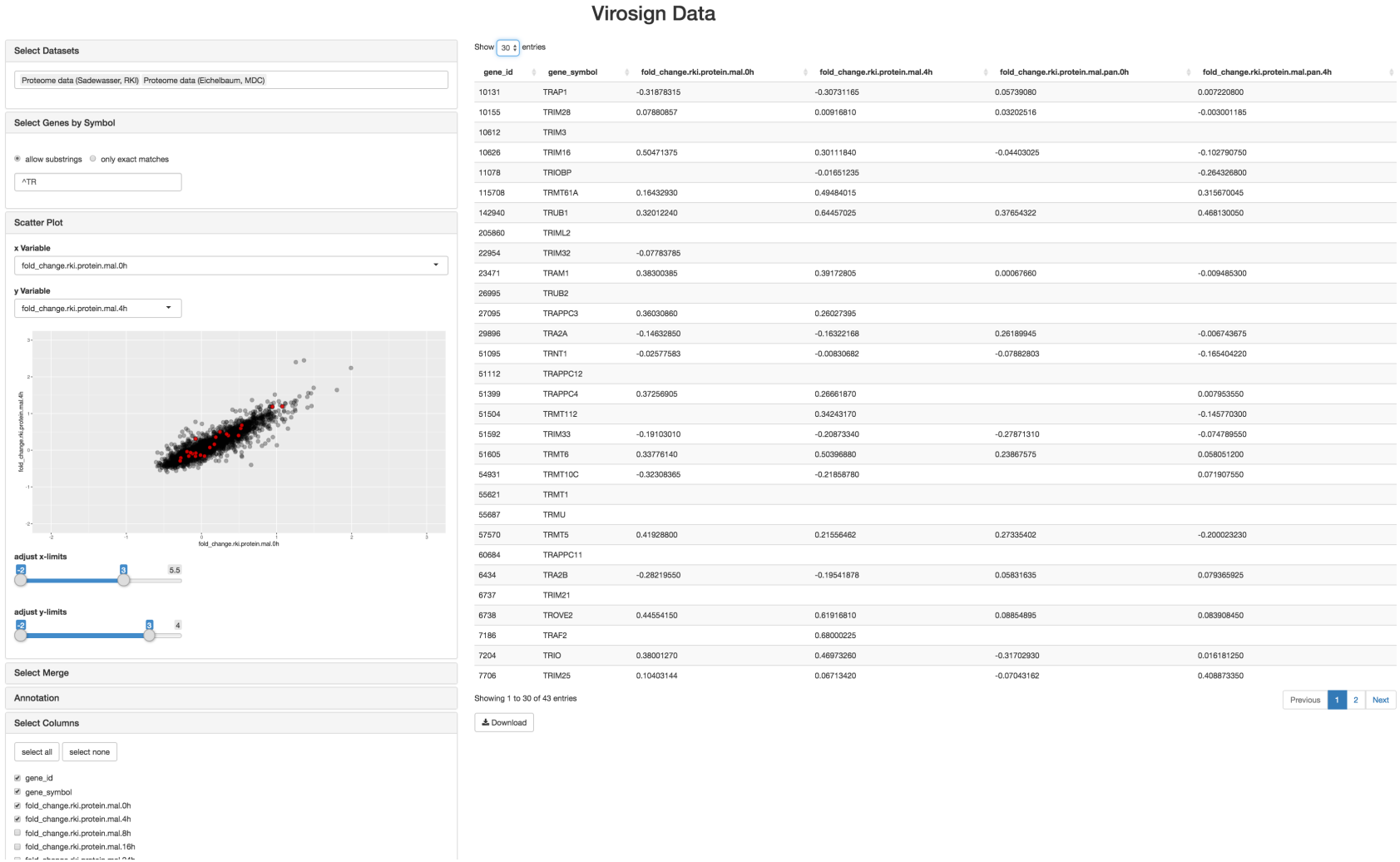

Fig. 3.4.: Screen shot of the ViroSign data interface. In the left panel, the data sets can be selected, genes can be filtered (highlighted in red in the scatter plot) and a data preview in form of a scatter plot of two selected columns is given. Furthermore, collapsible panels offer different modes of merging data sets, annotations and the selection of the columns to be shown in the table. On the right, the corresponding data table is shown and updates immediately when the settings panels are modified.

4 Mathematical modeling of the influenza A infection

> *After all it really is all of humanity that is under threat during a pandemic. [23]*
>
> — **Margaret Chan**
> (Director-General of WHO)

In this chapter, the intracellular mechanisms during influenza A infection are investigated using systems biology methods. Therefore, the reader is first introduced to the fundamentals of influenza biology. Subsequently, a mechanistic ODE model for the virus entry into the host cell is developed and refined through experimental measurements in living cells. To further investigate the transport of the viral genome to the host cell's nucleus, a spatial modeling strategy is applied. Combining the predictive power of both models, we obtain further insights into the early steps of influenza infection.

4.1 Introduction to influenza biology

The influenza A virus (IAV) belongs to the family of *Orthomyxoviridae* which are characterized by a segmented, negative sense, single-stranded RNA (ssRNA) genome. The nomenclature of IAV strains is based on the subtype of the two viral surface proteins hemagglutinin (HA) and neuraminidase (NA). The strain *H3N2*, for instance, refers to HA subtype 3, NA subtype 2. Furthermore, the strains are named by an isolate number, year and place of their first characterization, for instance, A/Panama/2007/1999 (H3N2) is a influenza A virus with the isolate number 2007 first characterized in 1999 in Panama with the subtype H3N2 [18].

4.1.1 The influenza A virus

The virion, as shown in Figure 4.1a, consists of a viral envelope containing glycoproteins HA and NA as well as the ion channel matrix protein 2 (M2). On the inside of the envelope

the viral matrix protein 1 (M1) stabilizes the capsid and is attached to the eight viral ribonucleoprotein (vRNPs) complexes coding for up to 11 viral proteins. These complexes are organized in rod-like structures comprising the ssRNA, the nucleoprotein (NP), small amounts of the nuclear export protein (NEP, also referred to as non-structural protein 2, NS2) and the polymerase complex (subunits PA, PB1, PB2) needed for transcription in the host cell (Figure 4.1b and c).

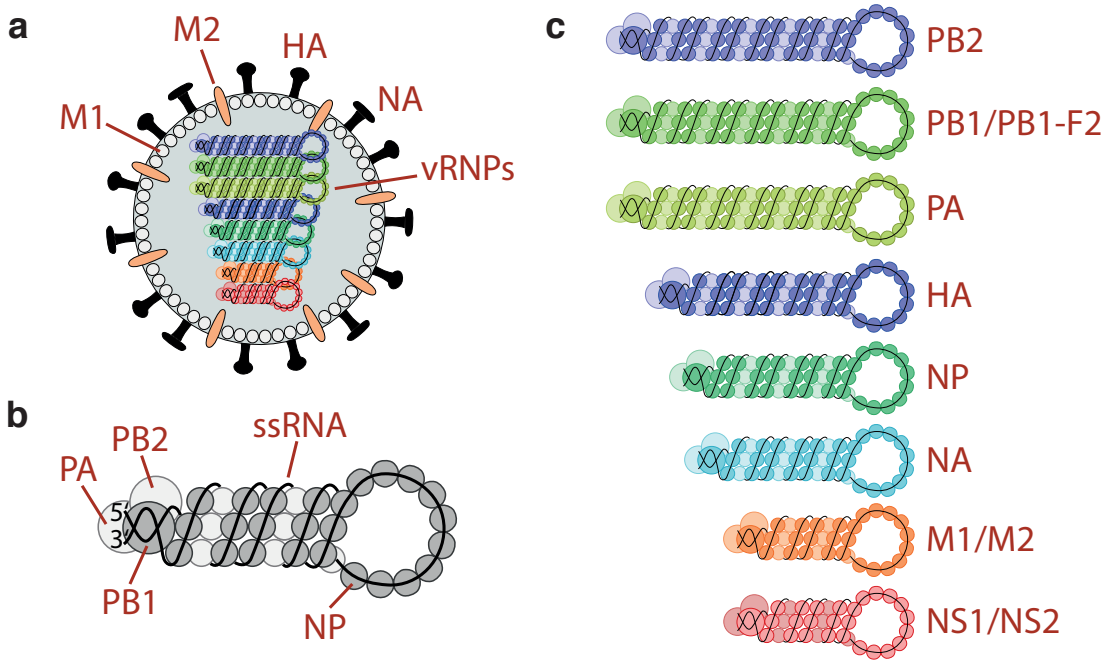

Fig. 4.1.: Structure and components of the influenza A virus. (a) Detailed schematic of the influenza A virus. The most abundant proteins are indicated. **(b)** Structure of an exemplary vRNP. Single-stranded RNA is wrapped around NP protein. The polymerase complex consisting of PA, PB1, PB2 is bundled with the RNP. **(c)** The eight genome segments and the proteins they code for. PB1-F2 is a splicing variant with unknown function.

4.1.2 Epidemiology of influenza A

The most recent H1N1 pandemic outbreak in 2009 killed more than $200\,000$ people within 12 month [39]. In addition, not only pandemic influenza constitutes a threat for human health. According to WHO, seasonal influenza strains are estimated to infect $5\,\%$ to $10\,\%$ of the world population every year [102]. Especially for the high-risk group, such as pregnant women, children and elderly people, seasonal infection can become life threatening causing up to $250\,000$ to $500\,000$ yearly deaths [102].

Antigenic variability

Influenza has two sources of genetic variability called *antigenic drift* and *antigenic shift*. While the *antigenic drift* is caused by simple point mutations during vRNA replication, *antigenic shift* occurs through a recombination of HA and NA subtypes from different virus strains mostly due to co-infection of a host organism with multiple strains. Genetic changes of seasonal influenza strains originate from *antigenic drift* while pandemic strains evolve due to *antigenic shift* Clements & Casani [33].

Vaccination

The first vaccines against seasonal influenza strains were invented in the 1940s. Due to the antigenic drift, these vaccines need to be adapted repeatedly. While the original vaccines were based on one specific inactivated strain, modern vaccines comprise three different strains that are predicted to be dominant in the upcoming influenza season. The selected vaccine strains are mainly grown in chicken eggs before they get inactivated. Only since 2016, the first cell-based vaccine was approved by the FDA which could reduce the offset of vaccine manufacturing dramatically [102]. Furthermore, so-called *universal vaccines* do not train the immune system for recognizing HA or NA antigens but the conserved *stalk* domain of the HA, NA or M2 protein [102].

Treatment

Two classes of antiviral treatment against influenza A have been approved by the drug authorities: (1) M2-blockers prevent the acidification of the virus and thus the release of the viral genome to the cytosol; (2) by inhibiting the viral NA protein, after budding, the progeny fails to dissociate from the infected host cell preventing a spread of infection Palese [148]. However, for both classes resistant strains have been reported necessitating the development of alternative antiviral approaches such as *small interfering RNA* (siRNA) Palese [148].

4.1.3 The replication cycle of influenza A

Virus entry

The replication cycle of IAV is depicted in Figure 4.2. After binding of HA to sialic acid receptors on the host cell's surface the virion gets internalized through endocytosis or

macropinocytosis [200]. During endosomal maturation, the lumen is acidified and the endosome travels along microtubules towards the microtubule organizing center (MTOC) in the peri-nuclear region [85]. In the late endosome, the pH in the lumen drops below a strain-specific threshold causing HA to undergo a conformational change. This triggers the fusion of the viral envelope with the endosomal membrane and leads to the escape of the vRNPs into the cytosol.

After the release, the vRNPs are transported to the nuclear membrane where they get imported based on their nuclear localization signals [37]. It is commonly assumed that the vRNPs is not actively transported, but diffuses to the nucleus [7]. Also, it is still under debate whether all eight vRNPs remain attached as one complex or if they dissociate shortly after endosomal fusion [127, 31].

Genome replication and synthesis of viral proteins

Being imported into the host cell's nucleus, the viral RNA polymerase, that was bundled in the vRNPs, starts to synthesize two positive-sense RNA templates: the vRNA that is used for the viral progeny and the complementary cRNA which (after stealing caps from host pre-mRNAs through cap-snatching [167]) can be used to produce viral proteins by the host's translation machinery. The surface proteins HA, NA and M2 are synthesized directly into the endoplasmatic reticulum (ER) and transported to the cell membrane by the Golgi apparatus while other viral proteins are produced in the cytoplasm of the host cell.

Transport and assembly of new virions

Rab11 positive vesicles transport the newly synthesized vRNPs to the cell membrane [6, 135, 3] where they assemble with the surface proteins, form a bud and are released to the extracellular space. As HA has a strong affinity for sialic acid receptors on the cell membrane, these bonds needs to be cleaved by means of the NA protein [18]. Therefore, NA-inhibitors like Oseltamivir (the active agent in *Tamiflu*) can been used to prevent the successful reproduction of IAV in the host organism [38].

The packaging strategy of the vRNPs was, for a long time, thought to be completely random [9, 86]. However, only recently Chou *et al.* [31] could show by colocalization analysis of data obtained by single-molecule fluorescence *in-situ* hybridization (smFISH) that selective packaging is more likely leading to virions with 8 unique vRNPs in place [32].

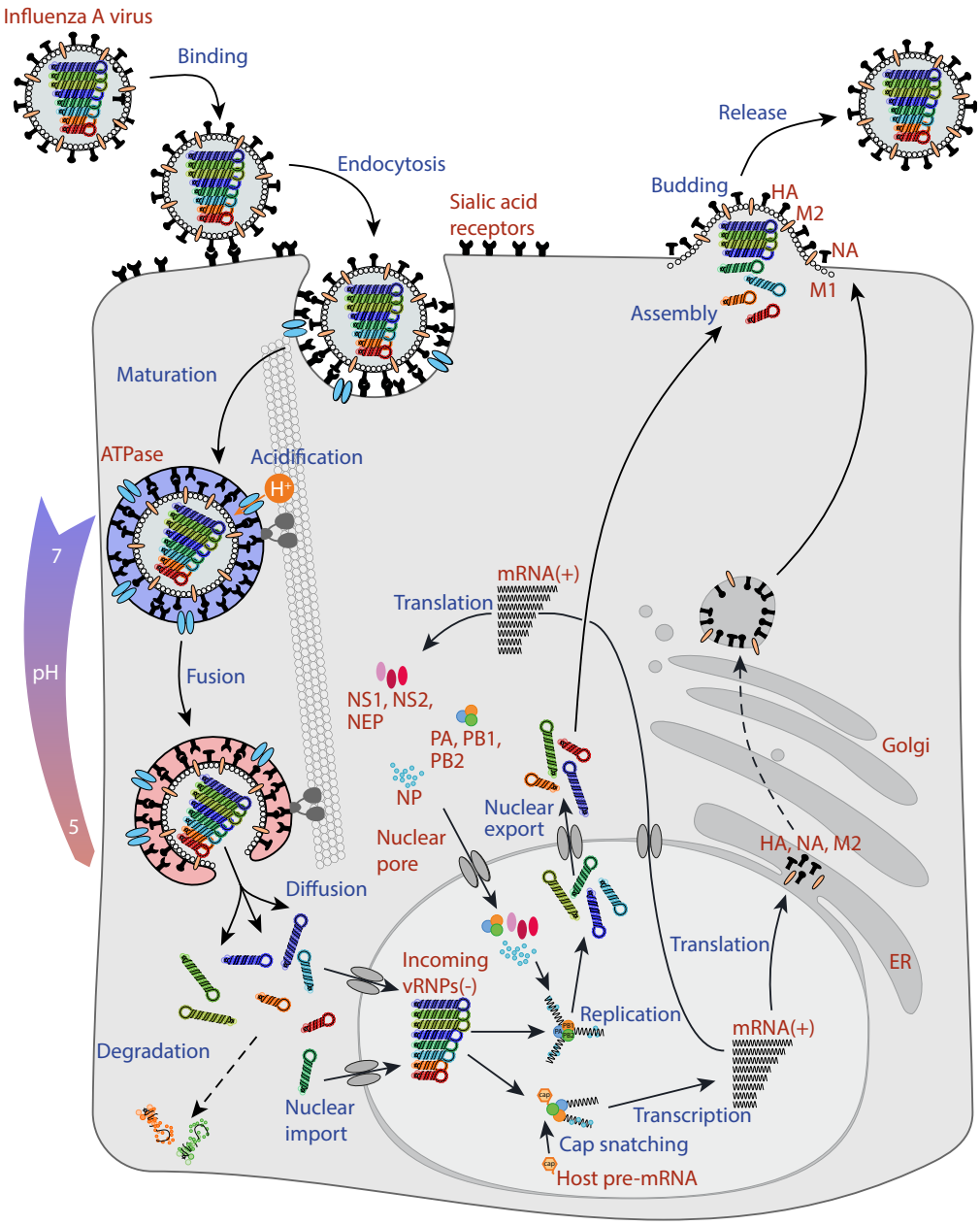

Fig. 4.2.: Replication cycle of the influenza A virus. A free virion binds to sialic acid receptors on the membrane of the host cell and is taken up via endocytosis. Endosomes travel along microtubules towards the MTOC. During maturation from early to late endosome the lumen of the endosome is acidified. When the pH threshold of the viral protein HA is reached, a conformational change causes the escape of the viral genome into the host cell's cytoplasm. Diffusive transport and active uptake through nuclear pores make the vRNPs enter into the nucleus host cell where the vRNA is cap-snatched with host pre-mRNA caps, transcribed and replicated. The viral proteins are synthesized and transported back to the nucleus (NP, M1, PA, PB1, PB2, NS1, NS2) or shuttled to the cell membrane via the Golgi (NA, HA, M2). vRNPs are formed and transported to the cell membrane where the viral progeny assembles and new virions bud and are released to the extracellular space.

For more exhaustive descriptions of the replication cycle I refer to Cheung & Poon [30], Bouvier & Palese [18], and Samji [173].

Virus-host interactions

In order to successfully infect an organism and replicate within a host cell, IAV relies on multiple interactions with the host cell. These interactions comprise the entry into the host cell through endocytosis, the escape from the endosome via membrane fusion but also interactions that suppress or evade the immune response. Furthermore, replication of viral RNA and synthesis of viral proteins strongly depends on the host machinery that is hijacked via cap-snatching. The immune response to IAV is comprehensively summarized by Kreijtz *et al.* [103]. Other interactions were reviewed in more detail e.g. by Madrahimov *et al.* [122].

4.2 Dynamic modeling of the influenza A infection

This part of the thesis is based on the publication

***Schelker, M.**, Mair, C. M., Jolmes, F., Welke, R.-W., Klipp, E., Herrmann, A., Flöttmann, M. & Sieben, C. „Viral RNA Degradation and Diffusion Act as a Bottleneck for the Influenza A Virus Infection Efficiency“.* PLOS Computational Biology *2016, **12**(10): e1005075*

All biological experiments have been conducted by CMM, FJ, RWW and CS. Model construction, simulation and analysis were performed by MF and MS.

As introduced in the previous section, the IAV infection is a multi-step process that critically depends on the viral protein HA and its sensitivity to altering pH values. Since endosomal transport towards the MTOC and acidification are concurrent, the distance that released vRNPs need to overcome depends on the location of membrane fusion and thereby on the pH-dependent conformational change of HA.

While endosomal transport is directed and actively driven by dynein and kinesin on microtubules [85], the movement of the viral genome after fusion is most likely passive and undirected [7]. Larger particles like the complexes of eight vRNP segments diffuse slower than their individual components due to their larger hydrodynamic radius. Nevertheless, packages comprising eight unique vRNP segments could be advantageous especially at a low *multiplicity of infection* (MOI) where a missing segment would cause a total failure

of the infection. Although the vRNA and NP are tightly packed within the vRNP, the genome remains sensitive to RNase digestion [214]. Therefore, it is unclear whether the diffusing vRNPs might be damaged due to cytosolic degradation as recently reported for the hepatitis C virus [15].

Here, we aimed to answer the question whether the entry of the virus and the transport of its genome to the host cell's nucleus might be limiting for the virus infection efficiency. More specifically, we wanted to identify crucial factors that may act as bottleneck during the infection. We therefore analyzed the virus entry using mathematical modeling combined with experimental data of the involved processes.

4.2.1 Existing models

The earliest models describing the population dynamics of mice infected with influenza were developed 40 years ago [113]. Since then, multiple models of IAV infection have been developed, representing the infection at different levels of detail. There are three main categories of influenza infection models: (1) epidemiological models analyzing and predicting the infection on a population scale; (2) multi-cellular models describing the spread of infection within the organism; and (3) intracellular models of IAV infection characterizing virus uptake and replication on a molecular level. An overview on modeling approaches of IAV infection is given in Beauchemin & Handel [11]. For a review on epidemiological models of IAV infection, I refer to Nsoesie *et al.* [144].

While the population scale is most evidently used for epidemiological risk prediction and vaccination strategies, more fine-grained models can help to identify drug targets or can be utilized to maximize the virus synthesis for cell culture-based vaccine production.

Here, we want to focus on the third category, which are models of intracellular processes involved during IAV infection. Sidorenko & Reichl [182] and later Heldt *et al.* [76] developed the first extensive intracellular models of IAV infection. However, due to the great level of detail, these models contain a great number of unknown parameters leading to a limited scope of reliable model predictions. In further work, intracellular and population scale models were combined to a multi-scale model [75] providing a more comprehensive model of virus infection.

Models of the intracellular dynamics of the Semliki Forest virus [41] and other acid-dependent enveloped viruses [40] have been suggested and include the pH sensitivity of the virus endosome fusion. However, as the pH dependency is not explicitly implemented, these models are not able to cope with multiple virus strains with altered pH thresholds.

4.2.2 An ODE model describing influenza A virus entry into the host cell

To investigate how pH sensitivity of HA influences the early IAV infection, existing models were not sufficient in their description of viral genome release from the endosome. Therefore, we built a relatively coarse-grained model of virus entry into the host cell with more detailed pH-dependent kinetics for the virus-endosome fusion.

Model structure

Through an iterative cycle of data generation and model refinement, we evolved to a first version of the model shown in Figure 4.3a. The model describes the following steps: (1) Virus-receptor complexes are internalized into endosomes; the viral genome escapes through (2) endosomal acidification and (3) a basal, meaning not pH-dependent, rate; vRNPs either diffuse towards the nucleus (4) or are degraded (5); vRNPs in the vicinity of the nucleus are imported (6). These six steps are described by the ODEs defined in (4.1–4.9):

$$\frac{\mathrm{d}[\mathrm{VirRec_{ex}}]}{\mathrm{d}t} = -k_{\mathrm{end}} \cdot [\mathrm{VirRec_{ex}}], \tag{4.1}$$

$$\frac{\mathrm{d}[\mathrm{VirRec_{end}}]}{\mathrm{d}t} = k_{\mathrm{end}} \cdot [\mathrm{VirRec_{ex}}] - k_{\mathrm{basal}} \cdot [\mathrm{VirRec_{end}}] - k_{\mathrm{fus}} \cdot \frac{[\mathrm{H^+_{end}}]^h \cdot [\mathrm{VirRec_{end}}]}{[\mathrm{H^+_{end}}]^h + {k_{\mathrm{H+}}}^h}, \tag{4.2}$$

$$\frac{\mathrm{d}[\mathrm{vRNP_{cyt}}]}{\mathrm{d}t} = k_{\mathrm{basal}} \cdot [\mathrm{VirRec_{end}}] + k_{\mathrm{fus}} \cdot \frac{[\mathrm{H^+_{end}}]^h \cdot [\mathrm{VirRec_{end}}]}{[\mathrm{H^+_{end}}]^h + {k_{\mathrm{H+}}}^h} - k_\tau \cdot [\mathrm{vRNP_{cyt}}] - k_{\mathrm{deg}} \cdot [\mathrm{vRNP_{cyt}}], \tag{4.3}$$

$$\frac{\mathrm{d}[\mathrm{vRNP_{cyt1}}]}{\mathrm{d}t} = k_\tau \cdot [\mathrm{vRNP_{cyt}}] - k_\tau \cdot [\mathrm{vRNP_{cyt1}}] - k_{\mathrm{deg}} \cdot [\mathrm{vRNP_{cyt1}}], \tag{4.4}$$

$$\frac{\mathrm{d}[\mathrm{vRNP_{cyt2}}]}{\mathrm{d}t} = k_\tau \cdot [\mathrm{vRNP_{cyt1}}] - k_\tau \cdot [\mathrm{vRNP_{cyt2}}] - k_{\mathrm{deg}} \cdot [\mathrm{vRNP_{cyt2}}], \tag{4.5}$$

$$\frac{\mathrm{d}[\mathrm{vRNP_{cyt3}}]}{\mathrm{d}t} = k_\tau \cdot [\mathrm{vRNP_{cyt2}}] - k_\tau \cdot [\mathrm{vRNP_{cyt3}}] - k_{\mathrm{deg}} \cdot [\mathrm{vRNP_{cyt3}}], \tag{4.6}$$

$$\frac{\mathrm{d}[\mathrm{vRNP_{cyt4}}]}{\mathrm{d}t} = k_\tau \cdot [\mathrm{vRNP_{cyt3}}] - k_\tau \cdot [\mathrm{vRNP_{cyt4}}] - k_{\mathrm{deg}} \cdot [\mathrm{vRNP_{cyt4}}], \tag{4.7}$$

$$\frac{\mathrm{d}[\mathrm{vRNP_{cyt5}}]}{\mathrm{d}t} = k_\tau \cdot [\mathrm{vRNP_{cyt4}}] - k_{\mathrm{deg}} \cdot [\mathrm{vRNP_{cyt5}}] - k_{\mathrm{imp}} \cdot \frac{[\mathrm{vRNP_{cyt5}}]}{k_{\mathrm{inhib}} + 1}, \tag{4.8}$$

$$\frac{\mathrm{d}[\mathrm{vRNP_{nuc}}]}{\mathrm{d}t} = k_{\mathrm{imp}} \cdot \frac{[\mathrm{vRNP_{cyt5}}]}{k_{\mathrm{inhib}} + 1}, \tag{4.9}$$

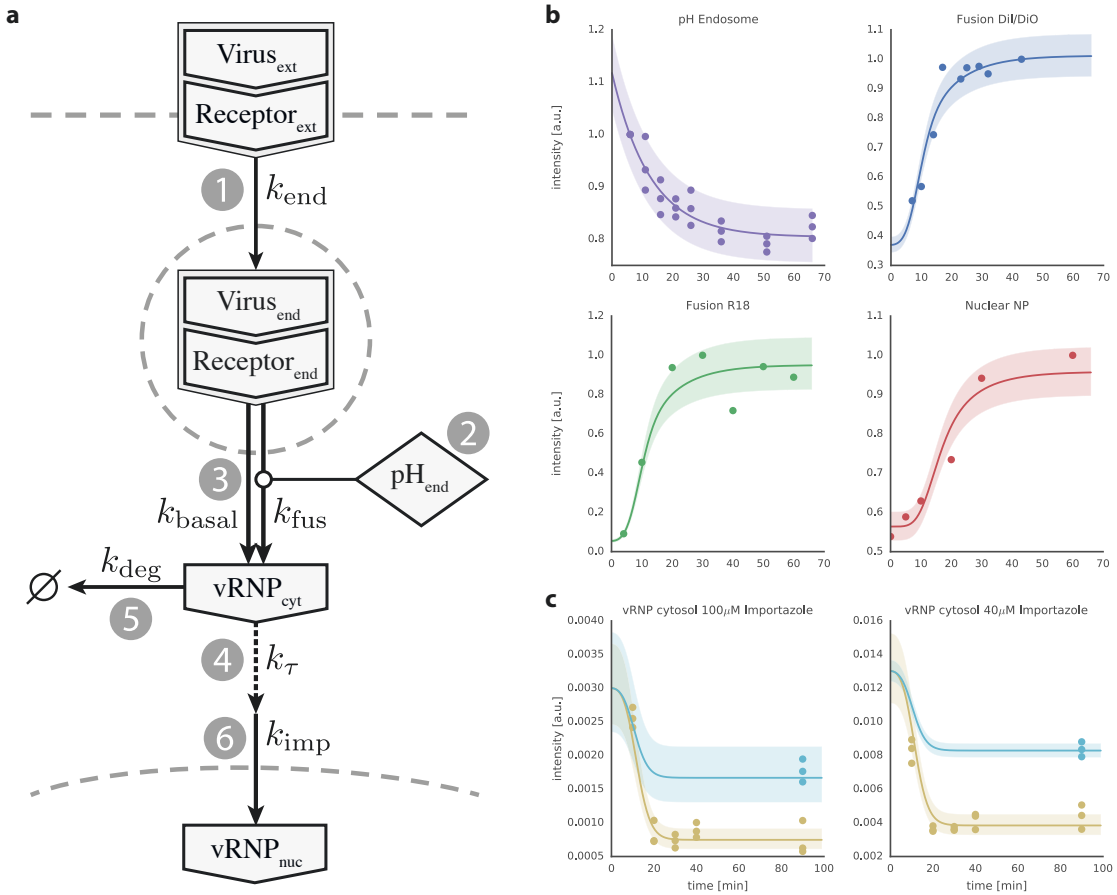

Fig. 4.3.: Model and data of the influenza A virus entry into the host cell. **(a)** Model structure of the ODE model. (1) Virus-receptor complexes are internalized into endosomes; (2) Endosomal maturation acidifies the lumen causing the release of the viral genome to the cytosol; (3) also pH-independent vRNP release is considered; (4) vRNPs diffuse towards the nucleus or (5) are degraded; (6) import into the host cell's nucleus completes the entry. **(b)** Measured data (dots) and fitted model dynamics (lines) of four observables. The shading represents the measurement error estimated based on a parametric error model. **(c)** Measured data and fitted model simulations for vRNP content in the cytosol (blue) and the whole cell (yellow). Cells were treated with different doses of importazole.

where [X] denotes the concentration of a model species X and the dynamics of endosome acidification are expressed as H^+ concentration:

$$[H^+_{end}] := 10^{-pH_{end}}. \tag{4.10}$$

This concentration is increasing proportionally to the ATPase activity (k_{ATPase}):

$$H^+_{end} = 10^{-(pH_{lb}+(pH_{ub}-pH_{lb})\cdot\exp(-k_{ATPase}\cdot t))}, \tag{4.11}$$

where pH_{ub} and pH_{lb} and are the upper (directly after endocytosis) and lower bounds (matured endosome) of the pH range in the endosome.

The delay τ caused by diffusion is implemented in the ODE model using a linear chain of $N = 5$ reactions (model species $[vRNP_{cyt1-5}]$) with simple mass action kinetics with the parameter

$$k_\tau = \frac{N}{\tau} \tag{4.12}$$

specifying the delay times [119].

Calibrating the model to experimental data

The model parameters were calibrated based on multiple sets of experimental data. Using flow cytometry, we could quantify the dynamics of the pH-level in the endosome. The percentage of fused endosomes was measured using fluorescence microscopy with two different fluorophores. The concentration of nuclear NP could be obtained as well by fluorescence microscopy leading to the following model observables:

$$pH_{obs} := -scale_{pH} \cdot \log_{10}([H^+_{end}]), \tag{4.13}$$

$$Fusion_{obs} := \log_{10}(offset_{Fus} + scale_{Fus} \cdot ([vRNP_{cyt,diff}] + [vRNP_{nuc}])), \tag{4.14}$$

$$NP_{obs} := \log_{10}(offset_{NP} + scale_{NP} \cdot [vRNP_{nuc}]). \tag{4.15}$$

Model fits and data are depicted in Figure 4.3b. At time point $t = 0$, viruses are bound to sialic acid receptors and internalization into endosomes is initiated by a temperature shift from $4\,°C$ to $37\,°C$. As the pH-level in the endosome decreases, after about $10\,min$, the threshold for the conformational change of HA is reached, the endosomal and viral membranes fuse as indicated by a steep increase in the both fusion measurements. After fusion, the released vRNPs travel towards the nucleus where they are imported leading

to an increase of the nuclear NP curve at around $15\,\mathrm{min}$. The time delay between fusion and nuclear import is modeled through a linear chain of reactions (4.4-4.8).

The pH-threshold at which fusion occurs can be determined by a dose-response experiment making use of a *fluorescence dequenching* (FDQ) assay. Viruses labeled with R18 at a selfquenching concentration are bound to sialic acid receptors on a ghost membrane from erythrocytes. Unbound viruses are removed by washing and the pH is decreased by addition of citric acid. The reduction in pH induces membrane fusion between viruses and the ghost membrane, which in turn leads to a reduction of the local R18 concentration and thereby to increase in fluorescence via dequenching. The fraction of fluorescence is calculated as

$$\mathrm{FDQ} = \frac{F(t) - F(0)}{F_{\max} - F(0)}, \tag{4.16}$$

where $F(0)$ and $F(t)$ represent the fluorescence intensity at $t = 0$ and at a given time t, respectively.

This experimental observation can be linked to the model by simulating the fraction of released vRNPs for different pH values

$$\mathrm{FDQ}_{\mathrm{obs}} := \frac{\mathrm{vRNP}_{\mathrm{cyt,diff}} + \mathrm{vRNP}_{\mathrm{nuc}}}{[\mathrm{VirRec}_{\mathrm{end}}]|_{\mathrm{t=0}}}. \tag{4.17}$$

Performing the FDQ experiment with different strains of IAV with an altered pH-sensitivity allows to investigate the model under various conditions.

Here, we make use of three IAV strains: (1) the A/X31 (H3N2) strain; (2) the recombinant WSN H3 wt strain which is generated from the A/WSN/1933 (H1N1) strain with the HA segment of the X31 strain; and (3) the recombinant WSN H3 mut strain, carrying the destabilizing double-mutation T212E-N216R in the HA protein of X31.

While the FDQ curves in Figure 4.4a for X31 and WSN H3 wt strains are very similar, the pH sensitivities for WSN H3 wild-type vs. the mutant strain are clearly distinct, as shown in Figure 4.4b. The FDQ of the X31 and WSN H3 wt strains is half-maximal at pH 5.6 while the WSN H3 mut strain has a steeper FDQ curve being half-maximal at pH 5.8.

As the genome release occurs during endosomal transport towards the MTOC, a shift in the pH-sensitivity to higher pH values, most likely, corresponds to a larger distance to the nucleus that needs to be overcome via diffusion. We therefore hypothesize that the mutant strain should show an altered infectivity, meaning less vRNPs arrive within a certain time, as compared to the wild-type. When simulating our ODE model for both conditions, however, we could not see any difference in the steady-state of NP in the nucleus. By contrast, measuring the infection efficiency in MDCK cells showed a $40\,\%$

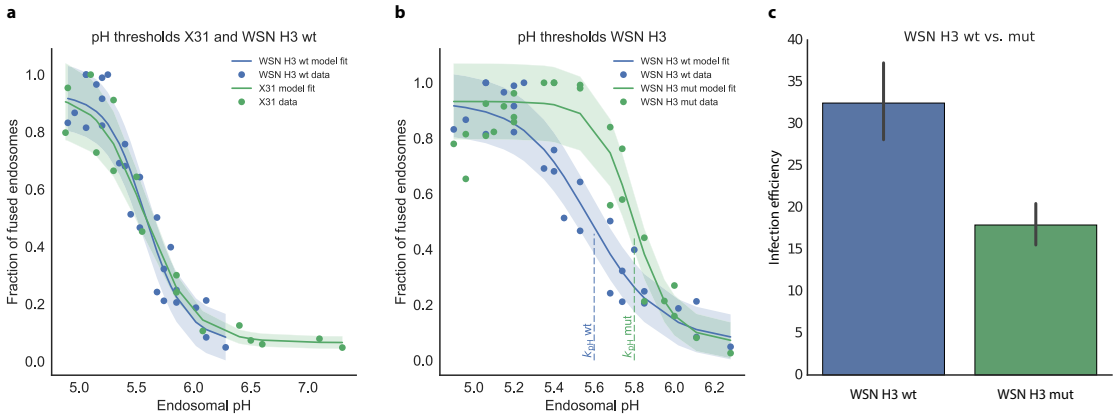

Fig. 4.4.: Fluorescence dequenching and infection efficiency for different strains of IAV. **(a)** Comparison of the pH sensitivity of the WSN H3 wt (green) and the X31 strain (blue). **(b)** Comparison of the WSN H3 wt (green) and the mutant (blue). Dots represent measured data, lines are the model simulations and the shading represents the measurement error estimated based on a parametric error model. **(c)** Comparison of the infection efficiency of WSN H3 wt and mut in MDCK cells measured by NP accumulation upon infection. Data show mean and SEM from four independent experiments.

to $50\,\%$ lower infectivity for the mutant strain, as depicted in Figure 4.4c. Thus, we concluded that there must be a missing reaction in our entry model, that reduces the amount of vRNPs over time during their diffusion through the cytosol.

Using R18-labeled X31 viruses, we could show that the virus-endosome fusion is taking place some micro-meters away from the nucleus (see Figure A.1 in the appendix). To investigate this crucial transport step, we introduced a spatial component that accounts for the diffusion of released vRNPs through the cytosol.

Parameter estimation and identifiability analysis

Parameter estimation was performed based on maximum likelihood estimation as described in Section 2.3.2. The model was implemented in the D2D software (Section 3.1) for MATLAB and a multi-start deterministic optimization strategy was applied. The model simulations of the observables are shown together with the experimental data in Figure 4.3b. The model trajectories for the best fit are depicted in Figure A.2 in the appendix.

Parameter identifiability was analyzed based on likelihood profiles (Section 2.3.3). As depicted in Figure 4.5, most parameters of interest are identifiable. Parameters characterizing the two WSN H3 strains show distinct parameter values for $k_{\mathrm{H}+}$ and

the Hill coefficient h. The maximum likelihood estimates and their likelihood-based confidence intervals are indicated in Table 4.1.

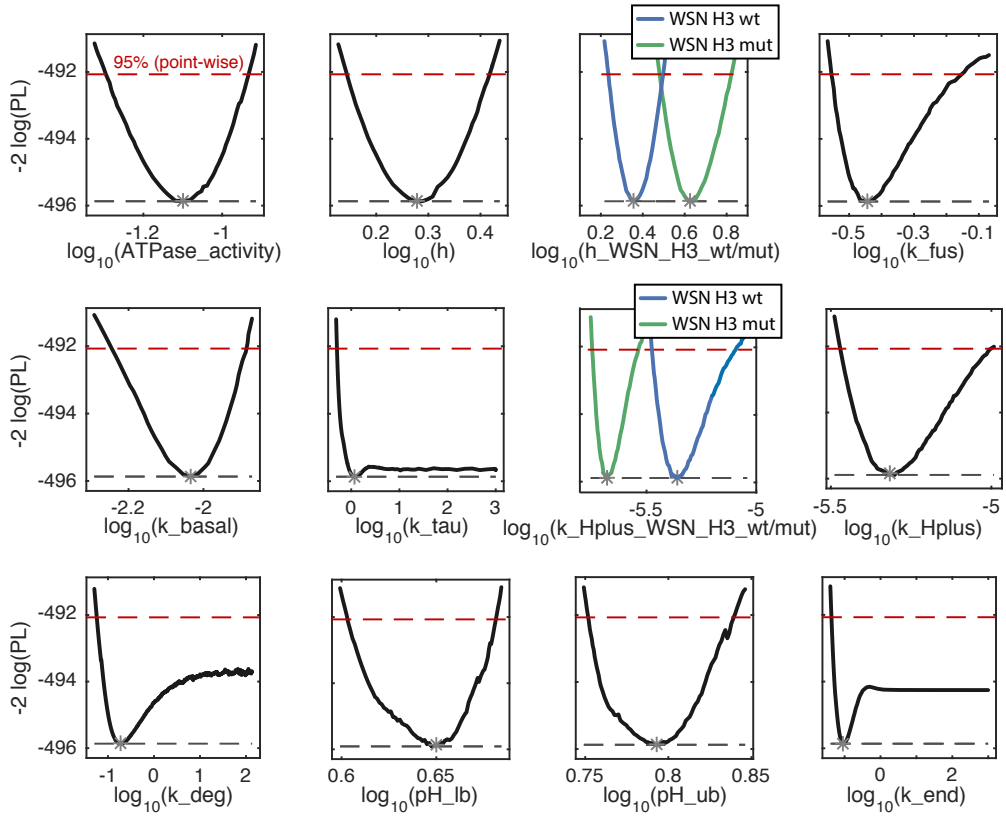

Fig. 4.5.: Profile likelihood analysis of selected model parameters. Lines represent the likelihood profile, point estimates are indicated by a gray asterisk. Red dashed lines indicates the $95\,\%$ point-wise confidence threshold. Parabola-shaped profiles indicate identifiable parameters. Parameter values are plotted on the log-scale.

4.2.3 A stochastic model of vRNP diffusion to the nucleus

After having identified diffusion of the viral genome from the point of cytosolic release to the nucleus as a critical step of the entry phase, we wanted to further investigate this process. Using a spatial-stochastic model, we simulated the transport from the point of virus-endosome fusion to the nucleus on a single molecule level. This allowed us to relate the fusion distance and delay to the efficiency of the IAV entry taking degradation of diffusing vRNPs, dissociation of vRNP packages and nuclear import into account.

We set up a reaction-diffusion model (see also Section 2.3.1) using the STEPS software [79]. Based on a given geometry and tetrahedral meshing, the model was simulated

Tab. 4.1.: Parameter names, optimal values, upper and lower bounds of $95\,\%$-confidence intervals and units are given for all kinetic model parameters. Confidence intervals were determined using the profile likelihood approach [162]. The plots of the likelihood profiles are depicted in Figure 4.5.

Parameter	$\hat{\theta}$	$\mathrm{conf}_{\mathrm{lb}}$	$\mathrm{conf}_{\mathrm{ub}}$	Unit
k_{ATPase}	7.94×10^{-2}	5.05×10^{-2}	1.17×10^{-1}	$[\mathrm{min}^{-1}]$
h	1.90	1.39	2.62	[1]
$h_{\mathrm{WSN\_H3\_mut}}$	4.25	3.02	6.67	[1]
$h_{\mathrm{WSN\_H3\_wt}}$	2.29	1.72	3.15	[1]
$k_{\mathrm{H+}}$	4.86×10^{-6}	3.38×10^{-6}	1.01×10^{-5}	$[\mathrm{mol} \cdot \mathrm{l}^{-1}]$
$k_{\mathrm{H+\_WSN\_H3\_mut}}$	2.10×10^{-6}	1.79×10^{-6}	2.91×10^{-6}	$[\mathrm{mol} \cdot \mathrm{l}^{-1}]$
$k_{\mathrm{H+\_WSN\_H3\_wt}}$	4.34×10^{-6}	3.34×10^{-6}	8.11×10^{-6}	$[\mathrm{mol} \cdot \mathrm{l}^{-1}]$
k_{basal}	9.24×10^{-3}	5.66×10^{-3}	1.31×10^{-2}	$[\mathrm{min}^{-1}]$
k_{deg}	1.89×10^{-1}	5.68×10^{-2}	∞	$[\mathrm{min}^{-1}]$
k_{end}	9.32×10^{-2}	4.28×10^{-2}	∞	$[\mathrm{min}^{-1}]$
k_{fus}	3.59×10^{-1}	2.79×10^{-1}	7.11×10^{-1}	$[\mathrm{min}^{-1}]$
k_{imp}	1.00×10^{4}	0.00	∞	$[\mathrm{min}^{-1}]$
$k_{\mathrm{inhib\_100}\mu\mathrm{M}}$	1.84×10^{-4}	0.00	∞	[1]
$k_{\mathrm{inhib\_40}\mu\mathrm{M}}$	9.89×10^{3}	0.00	∞	[1]
k_{τ}	1.21	5.04×10^{-1}	∞	$[\mathrm{min}^{-1}]$
$\mathrm{pH}_{\mathrm{lb}}$	4.46	4.01	4.80	$[\log_{10}(\mathrm{mol} \cdot \mathrm{l}^{-1})]$
$\mathrm{pH}_{\mathrm{ub}}$	6.20	5.65	6.91	$[\log_{10}(\mathrm{mol} \cdot \mathrm{l}^{-1})]$

with locally well-mixed reactions within the volume of each tetrahedron and stochastic transitions between tetrahedrons.

Performing spatio-temporal simulations of vRNP diffusion

The geometry of a generic MDCK cell was approximated by a cuboid as shown in Figure 4.6a. The volume is split into a tetrahedral mesh using tetgen [29] consisting of $25\,664$ tetrahedrons (thereof $22\,805$ in the cytosol and 2859 in the nucleus) with a maximal volume of $0.9\,\mathrm{fL}$ per tetrahedron.

The principle of stochastic transitions is shown on the example of a two-dimensional mesh in Figure 4.6b. The transition from tetrahedron k to the neighboring tetrahedron l takes place with the rate $d_{k,l}$ that is proportional to the coefficient of diffusion D. Furthermore, the surface $S_{k,l}$, which is the interface of both tretrahedra (in 2D represented by a line),

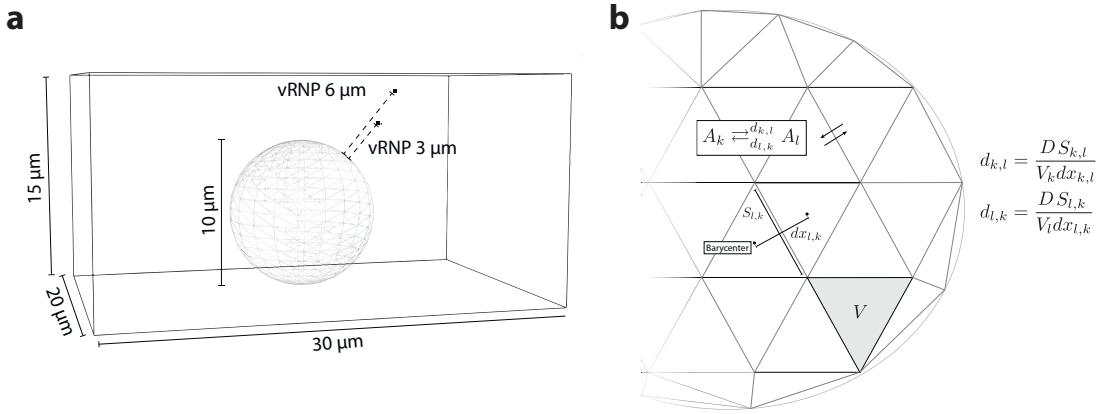

Fig. 4.6.: Cell-like geometry used for spatial simulations and schematic mesh grid in two dimensions. (a) The typical dimensions of a MDCK cell were adapted for a three-dimensional geometry based on a cuboid. The nucleus is represented by a sphere of $5\,\mu\mathrm{m}$ in radius situated in the lower third of the volume. **(b)** Schematic representation of a tetrahedral mesh of a circular structure in two dimensions. Measures introduced in the drawing are used to calculate the transition rates $d_{k,l}$ and $d_{l,k}$. D represents the diffusion coefficient.

the volume V_k (here the triangular surface) and the euclidean distance between both barycenters $dx_{k,l}$ adjust the transition rate.

In each volume, three possible well-mixed reactions were defined: (1) a vRNP package, containing eight vRNP segments, can dissociate into its segments with a rate constant k_{diss}; (2) both, vRNP packages and individual segments can be degraded during diffusion in the cytosol with a parameter k_{deg}; (3) vRNP packages and individual segments can bind to the nuclear membrane (k_{bind}) and are no longer subjected to degradation and dissociation.

Furthermore, the two initial conditions of the simulation, which are the distance to the nucleus d_{nuc} from the point of vRNP release and the number of released vRNP packages (corresponding to the MOI) need to be defined.

Exploring the parameter space

While the binding rate could be defined based on a measured value from literature [7] to $k_{\mathrm{bind}} = 0.056\,\mathrm{min}^{-1}$, the remaining parameters were not restricted by prior knowledge. Therefore, we analyzed the parameter space by performing two-dimensional scans for different pairs of parameters over several orders of magnitude. For each bin in Figure 4.7, a population of 1000 cells was simulated with parameter values indicated on the x and y-axis. The color-coding corresponds to the fraction of successful infections after $40\,\mathrm{min}$.

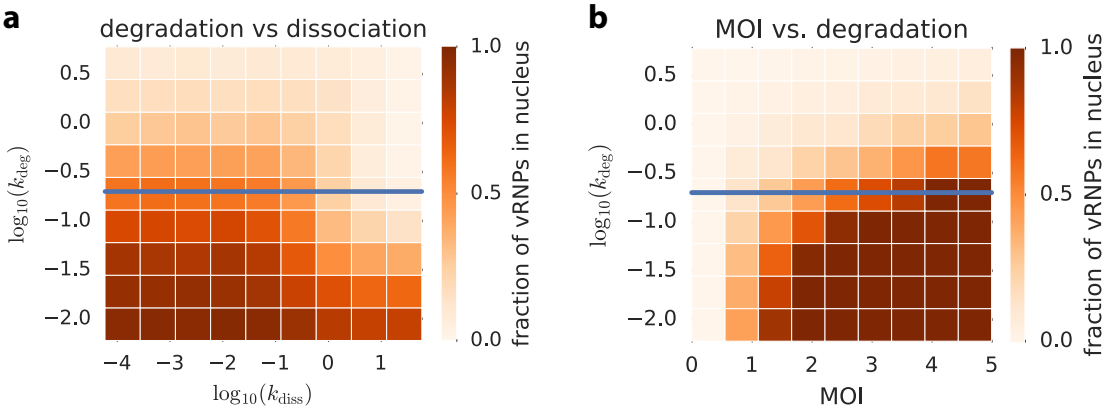

Fig. 4.7.: Dependency of the infection efficiency on the values of different model parameter. (a) Degradation vs. dissociation. High infection efficiencies were only observed for low dissociation and degradation values. For the experimentally determined degradation constant (blue horizontal line), low dissociation was favorable. **(b)** Degradation vs. multiplicity of infection. For the given degradation constant, only for high MOI values all cells were infected. For simulations shown in a) one vRNP was initially injected. In both panels, d_{nuc} was set to $3\,\mu\mathrm{m}$.

Here, *successful* is defined as arrival and binding of at least one complete set of vRNPs at the nuclear membrane.

The heat-map depicted in Figure 4.7a shows the influence of dissociation and degradation on the infection. For high dissociation and/or degradation constants, only a few out of the 1000 simulated cells got infected. Only for very small degradation constants ($k_{\mathrm{deg}} < 10^{-2}$) dissociated vRNP packages ($k_{\mathrm{diss}} > 10^{-1}$ were still able to reach infection efficiencies of $50\,\%$ and higher. The highest efficiencies were reached for both parameters being small ($k_{\mathrm{diss}} < 10^{-2}$ and $k_{\mathrm{deg}} < 10^{-1}$). Thus, despite the higher mean displacement of individual segments, our simulations clearly showed that the dissociation of complexes lowered the probability of a complete genome in the nucleus. The larger the dissociation constant in the model was set, the lower the percentage of complete genomes.

When looking at the impact of the initial amount of virus on the infection efficiency, as shown in Figure 4.7b, we found that for $\mathrm{MOI} > 1$ and $k_{\mathrm{deg}} < 10^{-1.5}$ an efficiency of more than $50\,\%$ was reached. When further increasing the MOI, also greater degradation constants could be compensated.

4.2.4 Combining both modeling approaches reveals vRNP removal in the cytosol

With both models at hand, we tried to narrow down the details of the influence of fusion distance to the infection efficiency. To quantify the rate constant of an assumed vRNP degradation reaction, we performed a RT-PCR experiment measuring the amount of vRNA of the HA segment (with specific primers for both, WSN H3 wt and mut) over time. We compared the cytosolic RNA levels after treating the cells with importazole, an inhibitor of nuclear import, with the total RNA levels as shown in Figure 4.3c. Fitting the ODE model parameters k_{deg} and k_{inhib} to the data revealed that there is a degradation or removal of vRNPs from the cytosol. Interestingly the inhibition of nuclear import was most prominent for the lower importazole dose ($40\,\mu\mathrm{M}$), most likely due to cell damage or the activation of other import pathways for the higher dose ($100\,\mu\mathrm{M}$) of the drug as reported by Kublun *et al.* [107].

Predicting the delay of the mutant strain

When simulating the amount of total and nuclear NP for both WSN H3 strains, we obtained the curves shown in Figure 4.8. The delay caused by the diffusion is denoted by τ_{eff} which corresponds to the half-maximal distance of the total and the nuclear NP curve for the wild-type. For the mutant strain, we saw an additional delay $\Delta\tau_{\mathrm{eff}}$ that was caused by the larger distance that the vRNPs need to diffuse in order to get imported to the nucleus.

Updating the spatial simulations

We used the estimated parameter value, which is $k_{\mathrm{deg}} = 0.19\,\mathrm{min}^{-1}$, for the spatial model, as depicted by the blue horizontal line in Figure 4.7. Next, we could translate the delay τ to the corresponding distance to the nucleus, as shown in Figure 4.9a for both, vRNP packages and complete sets of individual segments. With the delay estimated using the ODE model, we obtained a distance of $3\,\mu\mathrm{m}$ for the WSN H3 wt and $6\,\mu\mathrm{m}$ for the mutant strain. Simulating the spatial model again for these two values of d_{nuc}, we obtained curves shown in Figure 4.9b. While for large dissociation constants the infection efficiency was very low, for low dissociation constants, i.e. for stable vRNP packages, the infection efficiency was increased, up to $40\,\%$ for the wild-type and $20\,\%$ for the mutant strain.

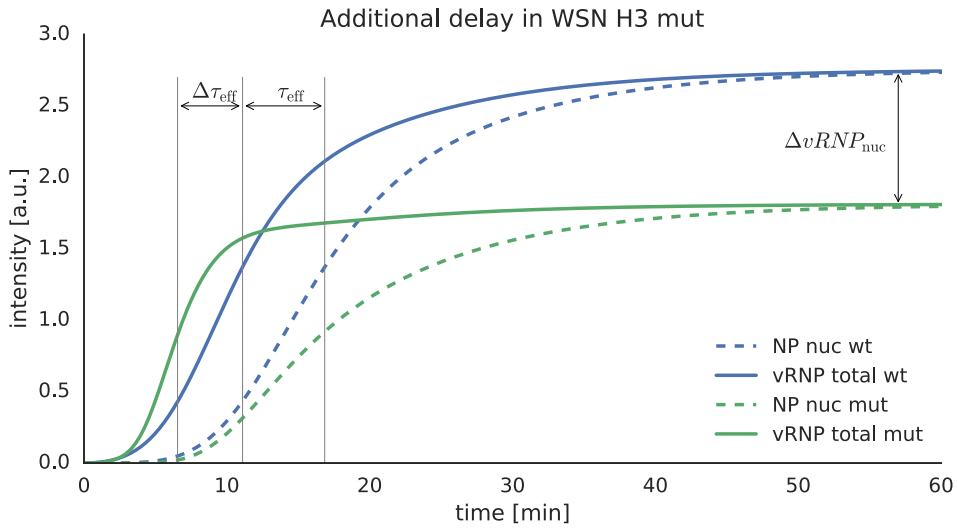

Fig. 4.8.: Model prediction of the amount of nuclear and total NP for the WSN H3 wt and mut strain. Solid lines represent the total NP concentrations (cytosol and nucleus), dashed lines show nuclear NP concentrations. The WSN H3 wt is depicted in blue, the mut strain in green. The delay between both strains, denoted as $\Delta\tau_{\text{eff}}$, was determined by simulating the ODE model with altered pH-sensitivity. τ_{eff} represents the delay between fusion and nuclear import. The difference in infection efficiency was determined by the concentration difference of nuclear NP after $60\,\text{min}$, as denoted by $\Delta\text{vRNP}_{\text{nuc}}$.

4.3 Discussion

Using a combination of a mechanistic intracellular model of the IAV entry into the host cell and a stochastic spatial model of the diffusion of released vRNP, we could show that stable packages are favorable to transport the genome into the nucleus.

We challenged the models with data from two IAV strains with altered pH sensitivity and found that an assumed degradation of vRNPs during diffusion could explain differences in the infectivity as observed experimentally. The existence of such a degradation reaction could be shown experimentally with RT-qPCR of the HA vRNA segment for cells with blocked nuclear import.

Taking both models together, we could translate the delay caused by diffusion to a distance and the simulation for both strains are consistent with the experimental observations.

The infection of a host cell with the influenza A virus represents a complex multi-step process with many viral and host factors being involved. The presented modeling approach captures the main characteristics of two different strains during the virus entry. Although the transport of the viral genome to the nucleus of the host cell is an

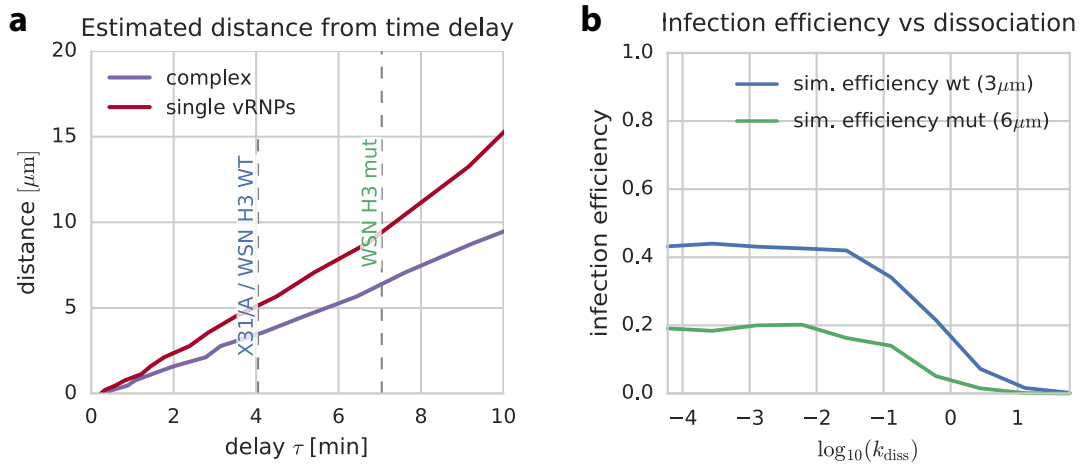

Fig. 4.9.: Estimating fusion distance and analyzing dependence of infection efficiency on dissociation. (a) Based on simulations of the spatial model, the delay is related to the fusion distance. Dashed lines indicate values for both IAV strains obtained from the ODE model. **(b)** Simulation of the infection efficiency for both strains. The different fusion distances result in a decreased infectivity of the mutant strain to $\sim 50\,\%$ of the wt value.

indispensable step during infection, differences in infectivity on the multi-cellular scale can have various sources. The investigated combination of passive transport and genome degradation provides one possible explanation for the observed phenomena.

Modern techniques such as single-molecule FISH with individually labeled genome segments could provide further insights into the critical steps of virus entry. Also, the spatial resolution and geometry of the model could be further refined to perform simulations in a more realistic setting.

Another possible limitation of our approach represents the lack of interaction with the immune system of the host in both models. The viral genome can be recognized through RIG-I that induces a type 1 interferon response [156]. However, it is not known whether the imported vRNPs are sufficient to trigger the innate immune response or only the newly synthesized ssRNA is detected by RIG-I [92].

5 Computational approaches for optimized treatment of cancer-associated health conditions

> *Cancer is not just a dividing cell. It's a complex disease: It invades, it metastasizes, it evades the immune system. [137]*
>
> — **Siddhartha Mukherjee**
> (Oncologist and author)

5.1 Introduction

Cancer is one of the world's major causes of death. According to WHO [188], in 2014 14.1 millions new cases of cancer occurred world wide and an estimated number of 8.2 million people died from cancer. While lung cancer still has the highest incidence and mortality rates among both, men and women ($13\,\%$), for women, breast cancer shows the highest incidence with more than $25\,\%$ of the total cases of cancer. Unlike for lung cancer, that is mainly caused by smoking, for other common types of cancer, the causes are often unknown and only risk factors such as overweight, obesity and familial predisposition can be identified. Besides prevention of cancer, for instance by dietary restrictions and increased physical activity, the early diagnosis can increase the chances of a success in therapy tremendously.

Treatment of cancer

The treatment of cancer depends on the specific indication of the patient. Typical approaches are surgical removal of tumors followed by radiation and/or chemotherapy. Other approaches involve specific antibodies or small-molecules that inhibit signaling pathways that are constantly activated in cancer cells [87].

Haematopoiesis and cancer

One prominent side effect of chemotherapy is anemia. Patients suffering from anemia are unable to produce sufficient amounts of red blood cells (erythrocytes) and therefore need to be treated either with blood transfusions or with injection of *erythropoiesis stimulating agents* (ESAs) like erythropoietin (Epo). The mode of action of Epo is based on the interaction of the hormone with the Epo receptor (EpoR) which induces differentiation of erythrocyte progenitor cells through activation of the JAK2-STAT5 signaling cascade. However, presence of the EpoR on the surface of various cancer cell types has been shown and anemia therapy using ESAs is therefore discussed controversially [78].

Immunotherapy as a novel treatment

While the human body is able to cope with various pathogens through the innate and adaptive immune response, for most types of cancer the immune system is incapable to successfully intervene and prevent the disease. Only recently, a new field of cancer treatments arose from this issue. Using immunotherapeutical drugs, the body's immune response is activated in order to fight cancer cells. The induced immune response is merely sufficient to combat an existing tumor but can render other therapies more efficient and reduce the side effects caused by the therapy [109].

Bridging the gap

Here, I want to present two projects that tackle cancer-associated issues with means of computational biology. The first aims to improve the current standard of care treatment for chemotherapy-associated anemia while the latter attempts to facilitate the quest for biomarkers for immunotherapy. In both cases, methods from computational biology bridge the gap between basic research and clinics and enable the selection of the best medication and dosing scheme on a patient-specific level.

5.2 Optimized treatment strategies for anemia patients based on a mechanistic multi-scale model

The following work emerged from joint project of experimentalists from the German Cancer Research Center in Heidelberg and theoreticians from University of Freiburg. Experiments and model simulations have been performed over the last 5 years in an iterative manner.

My contribution to this project consists of adapting the original model and data, incorporating new depletion data sets into the model, identifying model parameters and quantifying the uncertainties. Preliminary results of this project have been presented in the scope of my diploma thesis in 2012. However, as the project has made substantial progress since then, I want to present the final stage of it in this thesis.

To integrate this part into the bigger picture, in Section 5.2.6–5.2.7, I briefly present the continuation of the work which was mainly carried out by other people. A publication of the whole project with the title

Rodriguez-Gonzalez, A., ***Schelker, M.***, *Raue, A.*, *Steiert, B.*, *Böhm, M.*, *Salopiata, F.*, *Adlung, L.*, *Stepath, M.*, *Depner, S.*, *Wagner, M.-C.*, *Merkle, R.*, *Kramer, B. A.*, *Lattermann, S.*, *Wäsch, M.*, *Franke, A.*, *Klipp, E.*, *Wuchter, P.*, *Ho, A. D.*, *Lehmann, W. D.*, *Jarsch, M.*, *Schilling, M.*, *Timmer, J.* & *Klingmüller, U. „Mechanistic multiscale modelling enables personalized treatment“.* Manuscript in preparation *2017*

is currently in preparation. Furthermore, a patent titled

Rodriguez, A., *Schilling, M.*, *Klingmüller, U.*, *Raue, A.*, ***Schelker, M.***, *Timmer, J.*, *Jarsch, M.* & *Steiert, B. „Methods for the prediction of a personalized ESA-dose in the treatment of anemia“.* WO Patent App. PCT/EP2015/063,775 *2016*

was filed.

5.2.1 Introduction

Erythropoiesis

Renewal of red blood cells (erythrocytes) is achieved by differentiating *hematopoietic stem cells* (HSC) into erythrocytes. This process is tightly regulated by the cytokine

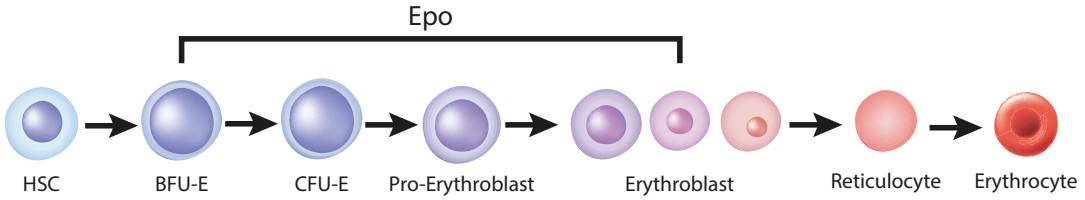

Fig. 5.1.: Schematic representation of the stages of erythropoiesis. Hematopoietic stem cells (HSC) differentiate to burst-forming unit erythroids (BFU-E) that express EpoR on their surface. Colony-forming unit erythroids (CFU-E) cells crucially depend on the presence of Epo for survival and differentiation to pro-erythroblasts. Erythroblasts eject their nucleus and differentiate to reticulocytes and eventually to red blood cells (erythrocytes). Figure adapted from Sankaran & Weiss [174].

erythropoietion (Epo). Based on the oxygen level in blood, the kidney regulates the release of Epo and thereby the production of red blood cells.

As depicted in Figure 5.1, there are several stages of maturation in erythropoiesis: pluripotent *hematopoietic stem cells* (HSC) in the bone marrow differentiate into *burst-forming unit erythroid cells* (BFU-E). While it has been shown that these early stages in erythropoiesis can exhibit EpoR on the cell surface [60], differentiation and cell survival does not critically depend on the presence of Epo at this stage [213]. BFU-E cells differentiate further into *colony forming unit erythroid cells* (CFU-E) that express substantial amounts of EpoR on the membrane. These cells can, upon stimulation with Epo, proliferate and differentiate into erythroblasts that expel their nucleus (reticulocyte) and are released into circulation where they mature into erythrocytes [174].

JAK2-STAT5 signaling

When Epo or a similar *erythropoiesis stimulating agent* (ESA) binds to a dimer of the corresponding Epo receptor (EpoR) on the cell surface, the receptor undergoes a conformational change resulting in the auto-phosphorylation of the *janus kinase 2* (JAK2) associated to the receptor. JAK2 in turn, phosphorylates the protein *signal transducer and activator of transcription 5* (STAT5) which dimerizes and translocates to the nucleus where it acts as a transcription activator for genes controlling cell fate decisions.

Anemia

In patients suffering from *chronic kidney disease* (CKD) or chemotherapy-associated anemia, the production of Epo and therefore of erythrocytes is reduced leading to an acute lack of red blood cells. In addition, the pro-inflammatory state of CKD causes

the release of other cytokines that inhibit maturation of blood progenitor cells [174]. This shortage of red blood cells is called anemia and may decrease the quality of life dramatically and contributes to morbidity [16]. Also, the decreased health-status may delay therapy of the primary diseases (such as cancer or CKD).

Treatment

Anemia is commonly treated by red blood cell transfusion or by administration of Epo to the patient. Variants of the recombinant human erythropoietin have been developed, such as C.E.R.A. and NESP, and offer a longer half-life *in vivo* while increasing the potency [120].

Adverse effects

Despite the general success in curing anemia using ESAs, the risk for hypertension and thrombotic events increases [192]. Furthermore, it was shown that cancer cells can also express low levels of the Epo receptor [78]. These receptors are potentially capable of activating the JAK2-STAT5 signaling pathway and therefore could increase survival and proliferation of cancer cells. These adverse effects are still discussed controversially [53].

A computational model of Epo-EpoR interactions

In the past, a model of Epo-EpoR dynamics has been developed and was successfully validated *in vitro* [12]. Here, this established model is utilized for characterizing different ESAs in both, mouse and human cell culture. The calibrated model enables us to determine not only the binding properties of new Epo variants but also can be utilized to estimate the number of binding sites of the cell type under investigation.

5.2.2 Building a mathematical model of ESA-EpoR interaction

Core model

The core model of Epo-EpoR interaction was originally developed by Becker *et al.* [12] and is depicted in Figure 5.2. It consists of a system of six ordinary differential equations

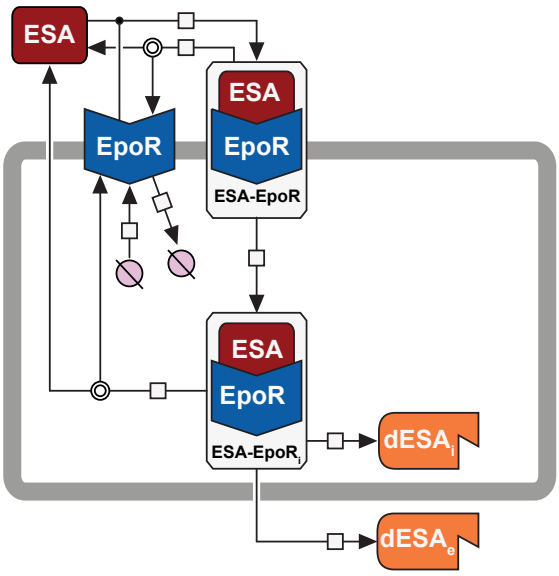

Fig. 5.2.: Model of EpoR receptor trafficking. A compatible ligand binds to the EpoR on the cell surface. Ligand-receptor complexes are internalized and degraded. A constant turnover of EpoR ensures responsiveness of the cell to ESA stimuli on a broad dynamic range. Figure adapted from Rodriguez-Gonzalez *et al.* [170].

comprising seven kinetic parameters and additional six initial concentrations of the model species.

$$\frac{\mathrm{d}[\mathrm{ESA}]}{\mathrm{d}t} = -[\mathrm{ESA}]\cdot[\mathrm{EpoR}]\cdot k_{\mathrm{on}} + [\mathrm{ESA\text{-}EpoR}]\cdot k_{\mathrm{off}} + [\mathrm{ESA\text{-}EpoR_i}]\cdot k_{\mathrm{ex}} \tag{5.1}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{EpoR}]}{\mathrm{d}t} &= -[\mathrm{ESA}]\cdot[\mathrm{EpoR}]\cdot k_{\mathrm{on}} + [\mathrm{ESA\text{-}EpoR}]\cdot k_{\mathrm{off}} + \mathrm{ESA}_{\mathrm{bind}}\cdot k_{\mathrm{t}} \\ &\quad - [\mathrm{EpoR}]\cdot k_{\mathrm{t}} + [\mathrm{ESA\text{-}EpoR_i}]\cdot k_{\mathrm{ex}}\end{aligned} \tag{5.2}$$

$$\frac{\mathrm{d}[\mathrm{ESA\text{-}EpoR}]}{\mathrm{d}t} = [\mathrm{ESA}]\cdot[\mathrm{EpoR}]\cdot k_{\mathrm{on}} - [\mathrm{ESA\text{-}EpoR}]\cdot k_{\mathrm{off}} - [\mathrm{ESA\text{-}EpoR}]\cdot k_{\mathrm{e}} \tag{5.3}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{ESA\text{-}EpoR_i}]}{\mathrm{d}t} &= [\mathrm{ESA\text{-}EpoR}]\cdot k_{\mathrm{e}} - [\mathrm{ESA\text{-}EpoR_i}]\cdot k_{\mathrm{ex}} - [\mathrm{ESA\text{-}EpoR_i}]\cdot k_{\mathrm{di}} \\ &\quad - [\mathrm{ESA\text{-}EpoR_i}]\cdot k_{\mathrm{de}}\end{aligned} \tag{5.4}$$

$$\frac{\mathrm{d}[\mathrm{dESA_i}]}{\mathrm{d}t} = [\mathrm{ESA\text{-}EpoR_i}]\cdot k_{\mathrm{di}} \tag{5.5}$$

$$\frac{\mathrm{d}[\mathrm{dESA_e}]}{\mathrm{d}t} = [\mathrm{ESA\text{-}EpoR_i}]\cdot k_{\mathrm{de}} \tag{5.6}$$

The reactions in (5.1)–(5.6) describe the biological processes of association and dissociation of the ligand Epo (more generally any ESA) to the receptor (EpoR), the ligand-receptor complex' internalization, dissociation as well as the degradation and export of ligand inside the cell. Moreover, the receptors on the cell surface show a

constant turn-over, as modeled by a production (proportional to the initial number of binding sites $\mathrm{ESA_{bind}}$) and degradation term.

Auxiliary model for streptavidin kinetics

To characterize the turnover rate k_t, a variant of the core model was implemented by Becker *et al.* [12] that describes the interaction dynamics of streptavidin (SAv) and its binding to a SAv-binding peptide (SBP)-tagged EpoR. While the binding properties of SAv to the tagged EpoR and the internalization rates might be very different, the constant turnover of unoccupied EpoR remains unchanged. Therefore, the parameters of both, the core and the auxiliary model could be simultaneously estimated.

A sub-model for binding kinetics

A small static model was created for characterizing the ligand-receptor interactions as measured by the dose-response experiments. In the original work [12], a one-site saturation regression (Michaelis-Menten kinetics) was utilized to fit the data

$$[\mathrm{ESA_{spec\_bound}}] = \frac{\mathrm{ESA_{bind}} \cdot [\mathrm{ESA_{free}}]}{[\mathrm{ESA_{free}}] + K_\mathrm{D}} \tag{5.7}$$

where $\mathrm{ESA_{bind}}$ denotes the number of ESA binding sites and K_D is the dissociation constant which is linked to the association and dissociation rate constants k_on and k_off by the relation

$$K_\mathrm{D} = \frac{k_\mathrm{off}}{k_\mathrm{on}}. \tag{5.8}$$

Experimental data

The original data from Becker *et al.* [12] consist of time course data of Epo alfa and SAv as well as dose-response data of Epo alfa and SAv. The data was acquired by measuring the radioactivity of ligands labeled with the instable isotope $^{125}\mathrm{I}$. The experiments were performed in BaF3-HA-mEpoR and BaF3-SBP-EpoR cells respectively.

With this technique, they could monitor the dynamic changes in radioactivity of three quantities, (1) extracellular Epo ($\mathrm{ESA_{ext}}$), (2) membrane bound Epo ($\mathrm{ESA_{mem}}$), and (3)

internalized Epo ($\mathrm{ESA_{int}}$). These could be mapped to the model states by the following relations:

$$[\mathrm{ESA_{ext}}] := [\mathrm{ESA}] + [\mathrm{dESA_e}] \tag{5.9}$$

$$[\mathrm{ESA_{mem}}] := [\mathrm{ESA\text{-}EpoR}] \tag{5.10}$$

$$[\mathrm{ESA_{int}}] := [\mathrm{ESA\text{-}EpoR_i}] + [\mathrm{dESA_i}] \tag{5.11}$$

To quantify the binding properties of Epo alfa and SAv, a binding assay was performed. For this purpose, BaF3-HA-mEpoR and BaF3-SBP-mEpoR cells were incubated for $4\,\mathrm{h}$ at $21\,^{\circ}\mathrm{C}$ with different doses of the ^{125}I-labeled ligand.

Parameter estimation and identifiability

The two parameters from the sub-model describing the binding kinetics, i.e. $\mathrm{ESA_{bind}}$ and K_{D}, were plugged into the core model of Epo-EpoR interaction and its variant for SAv respectively. The remaining model parameters describing association, internalization and degradation of Epo-EpoR complexes were estimated simultaneously for both, the core and the auxiliary model.

Identifiability analysis based on non-parametric bootstrapping [77] and on the profile likelihood method (see Section 2.3.3) was performed and all parameters could be estimated within tight confidence bounds.

Model predictions and validation

Using the calibrated model, they generated two hypotheses: (1) Epo is quickly depleted from the medium due to internalization of Epo-EpoR complexes followed by degradation of the ligand, and (2) Epo receptors on the cell surface recover rapidly after stimulation with Epo.

These predictions were then validated experimentally. Therefore, BaF3-HA-EpoR were treated with $5\,\mathrm{U/ml}$ of Epo alfa and the cultured medium was used to stimulate a second cell pool. As a read-out for the presence of Epo in the medium, phosphorylated EpoR (pEpoR) was monitored over $300\,\mathrm{min}$ in both cell pools. Quantitative immunoblotting confirmed the prediction and showed lower pEpoR values for the pool stimulated using the cultured medium as compared to the control.

Also the second prediction could be confirmed by quantitative immunoblotting. After stimulating BaF3-HA-EpoR cells with $5\,\mathrm{U/ml}$ of Epo alfa, pEpoR level returned to the

basal level within $60\,\mathrm{min}$ to $120\,\mathrm{min}$ after stimulation. When stimulating these cells again with an excess of ligand ($50\,\mathrm{U/ml}$), the phosphorylated EpoR could still reach the same levels as in the initial activation validating the hypothesis of rapid replenishment of EpoR on the cell surface.

5.2.3 Extending the model for multiple cell types and ESAs

Although the model was successfully used to describe the Epo alfa and SAv interactions at the receptor level in BaF3-HA-EpoR cells, several modifications needed to be implemented in order to include novel data for other ESAs and cell types.

Correcting the time course data for experimental delay

The existing time course data were adjusted by shifting the recorded time points by a time shift τ. This was necessary due to limitations in the experimental setup that had not been taken into account in the original work.

The exact value of τ is determined by a series of model fits with increasing values of τ where in each step, the corresponding time points were shifted to $t \to t + \tau$. Using this approach, a profile for each time course experiment is generated. The best value for τ was found to be $\tau = 0.82\,\mathrm{min}$ for Epo alfa replicate 1, $\tau = 0.76\,\mathrm{min}$ for Epo alfa replicate 2 and $\tau = 2.19\,\mathrm{min}$ for the SAv time course data.

Fitting all sub-models simultaneously

While in the original setup the sub-model for ESA binding was fitted separately and the resulting parameters K_D and $\mathrm{ESA_{bind}}$ were then plugged into the dynamic model, one could also fit both models simultaneously thereby making sure that uncertainties in the estimation are correctly propagated. However, in the process it turned out that the affinity constants obtained using the binding assay are not fully compatible with those controlling the dynamics in the ESA-EpoR model. This is most likely due to the difference in temperature the two experiments were performed at ($21\,^\circ\mathrm{C}$ vs. $37\,^\circ\mathrm{C}$) and was handled by allowing for a individual K_D parameter for both settings.

Using Hill-kinetics for the binding model

Another modification was necessary to link the dynamic core model and the static binding model. In the original work the binding process was described as Michaelis-Menten type saturation kinetics. However, the shape of the dose-response data for SAv could not be captured accurately with this assumption. We therefore substituted the one-site saturation by the more general Hill kinetics

$$[\mathrm{ESA}_{\mathrm{spec\_bound}}] = \frac{\mathrm{ESA}_{\mathrm{bind}} \cdot ([\mathrm{ESA}_{\mathrm{free}}])^n}{([\mathrm{ESA}_{\mathrm{free}}])^n + (K_{\mathrm{D}})^n} \tag{5.12}$$

and estimated the Hill coefficient n together with the other parameters. Also, as both the dependent and the independent variable in (5.12) represent protein concentrations, the model is fitted on a log-log scale to account for the log-normally distributed uncertainties of the data [104].

Incorporating individual data points instead of mean and standard error

As described more extensively in the methods chapter (see Section 2.3.2), preprocessing of experimental data by taking mean and standard error of the mean over several replicates can be inaccurate when the number of replicates is low. An alternative approach to quantify measurement uncertainties is to setup a parametric error model that is calibrated simultaneously with the model parameters.

Parameter estimation on the log-scale

As the uncertainties of protein measurements are most likely described by a log-normal distribution[104, 164], all model quantities and data points describing protein concentrations are transformed to the log-scale. For the time course data of Becker *et al.* [12], a scaling and offset parameter is introduced to account for the relative nature of the experiment and possible background radiation. After these adjustments, the observables defined in (5.9)–(5.11) are replaced by

$$[\mathrm{ESA}_{\mathrm{ext}}] = \log_{10}\left(\mathrm{offset}_j + \mathrm{scale}_j \cdot ([\mathrm{ESA}] + [\mathrm{dESA}_{\mathrm{e}}])\right), \tag{5.13}$$

$$[\mathrm{ESA}_{\mathrm{mem}}] = \log_{10}\left(\mathrm{offset}_j + \mathrm{scale}_j \cdot [\text{ESA-EpoR}]\right), \tag{5.14}$$

$$[\mathrm{ESA}_{\mathrm{int}}] = \log_{10}\left(\mathrm{offset}_j + \mathrm{scale}_j \cdot ([\text{ESA-EpoR}_{\mathrm{i}}] + [\mathrm{dESA}_{\mathrm{i}}])\right), \tag{5.15}$$

where the index j indicates that the parameters can be experiment-specific.

Tab. 5.1.: Experimental combinations of different ESAs and cell lines. For each combination the respective number of data points is indicated.

	Epo alfa	Epo beta	NESP	C.E.R.A.
BaF3 parental	40	40	12	26
BaF3-HA-mEpoR	54	65	0	0
BaF3-HA-hEpoR	60	74	46	70
hCD34+	0	17	0	0
hCFU-E	0	28	0	0
H838	83	184	0	0
H838-HA-hEpoR	42	42	144	24
H1944	0	42	0	0
H1299	0	63	0	0
A549	0	168	0	0

5.2.4 Incorporating ligand depletion experiments for various cell lines and ESAs

In this follow-up project, the original model was utilized and new data for different cell lines and ESAs were added. Therefore, we assume that most kinetic parameters remain unchanged. Only the constants describing the binding kinetics (i.e. k_{on} and k_{off}) and the number of binding sites on each cell (i.e. $\mathrm{ESA}_{\mathrm{bind}}$) are thought to be ESA and cell-specific properties.

The radio-labeling experiments, as performed by Becker *et al.* [12], were very complex and could not be adapted for other cell lines and ESAs. Therefore, a new experimental platform was established that measures the depletion of ESA in the supernatant of cells based on enzyme-linked immunosorbent assay (ELISA). This depletion data consists of a panel of four ESAs and 13 cell types leading to 77 time course measurements with a total of 1405 data points. In Table 5.1 the number of data points is indicated for each combination of ESAs and cell type.

Making depletion experiments comparable

As we included data from various cell types treated with different doses of several ESAs, the experimental setup for each depletion experiment could be very different in terms of the used volume of medium V and the number of cells N contained in each well. Therefore, a standardized volume $\hat{V}$ and cell number $\hat{N}$ were defined

$$\hat{V} = 0.1\,\mathrm{ml}, \qquad \hat{N} = 1 \times 10^6 \tag{5.16}$$

and the parameter $\mathrm{ESA}_{\mathrm{bind}}$ representing the number of ESA binding sites on a per cell basis was converted accordingly.

Moreover, for cell lines that express EpoR through retro-viral transfection, the transfection efficiency $\varphi \leq 1$ needs to be taken into account and the number of cells is corrected by

$$N = \varphi \cdot \tilde{N}. \tag{5.17}$$

Mapping the depletion data to the model species

The ELISA experiment quantifies the concentration of ESAs in the supernatant. Due to the specificity of the used antibodies, the observation can be directly linked to the corresponding model species:

$$[\mathrm{ESA}_{\mathrm{depl}}] = \log_{10}\left(\mathrm{offset}_j + \mathrm{scale}_j \cdot [\mathrm{ESA}]\right). \tag{5.18}$$

Again, we use only relative measurements and thus need to introduce parameters accounting for the scaling and possible background. Also the log-transformation is applied respect the underlying error distribution. For the error model, an absolute error on the log-scale was assumed

$$f_{\sigma_j} = \mathrm{sd}_j \tag{5.19}$$

which corresponds to a relative error in the non-log space.

Linking relative data to the absolute concentration scale

For both types of time course data, the radio-labeling and the ELISA, we deal with relative data that is mapped to absolute concentrations in the model. To remove the structural non-identifiability of the scale_j parameters, an additional data point reflecting the amount of the pipetted ESA dose is taken into account.

$$[\mathrm{ESA}_{\mathrm{depl}}]|_{t=0} = \log_{10}\left([\mathrm{ESA}]|_{t=0}\right). \tag{5.20}$$

As there is no actual control for the uncertainty of this value, we fixed the error model for this observable to a relative standard deviation of $10\,\%$

$$\mathrm{sd}_j = \log_{10}(1.1). \tag{5.21}$$

Model calibration and identifiability analysis

In order to utilize the model as a tool for characterizing different ESAs and cell lines, the model calibration was split into two separate fitting setups. In the first setting, the model was calibrated to all data obtained using the BaF3 cell line expressing the murine or human EpoR or its negative control that were stimulated with one of the four different ESAs. While the receptor-ligand affinities are ESA-specific and the number of binding sites, as described by the parameter $\mathrm{ESA}_{\mathrm{bind}}$, are cell line-specific, all other model parameters were assumed to be invariant under these different experimental conditions. In the second setting, all model parameters were fixed to the optimal value for the BaF3 cells and only the cell line-specific parameters were fitted to the data.

For parameter estimation and identifiability analysis, the D2D software was used, which is described in more detail in Section 3.1. The best fit was determined by minimizing the negative log-likelihood using the deterministic optimization algorithm `lsqnonlin` in MATLAB (R2016a, The MathWorks Inc., Natick, MA) with randomly sampled initial values (multi-start strategy, see also Section 2.3.2 and Section 3.1.2). The identifiability was determined by calculating the profile likelihood (see Section 2.3.3). This method also provides the corresponding confidence intervals. The corresponding likelihood profiles are shown in Figure B.6 and Figure B.7, respectively for both settings. The full table of parameter estimates from both settings and their confidence intervals are given in Table B.1 in the appendix.

5.2.5 Results

The data of Becker *et al.* [12] revisited

In Figure 5.3, the experimental data by Becker *et al.* [12] and the corresponding model simulations of the best fit are depicted. The model observables show the concentration of Epo alfa and SAv in medium, on the cell surface and within the cells. Additionally, the binding kinetics are shown by plotting the specifically bound ligand against the free ligand.

As the experimental data was obtained in a radio-labeling experiment, the observable for Epo in medium, as defined in (5.9), also includes the degraded, non-functional form of the ligand. Therefore, instead of a pure depletion, an increase of $[\mathrm{ESA}_{\mathrm{ext}}]$ after $\sim 120\,\mathrm{min}$ is observed. By contrast, for the dynamics of SAv, due to the shorter measurement time and the lower affinity to the receptor, no changes are seen in the medium and the ligand concentrations on the membrane and within the cells increase only slowly.

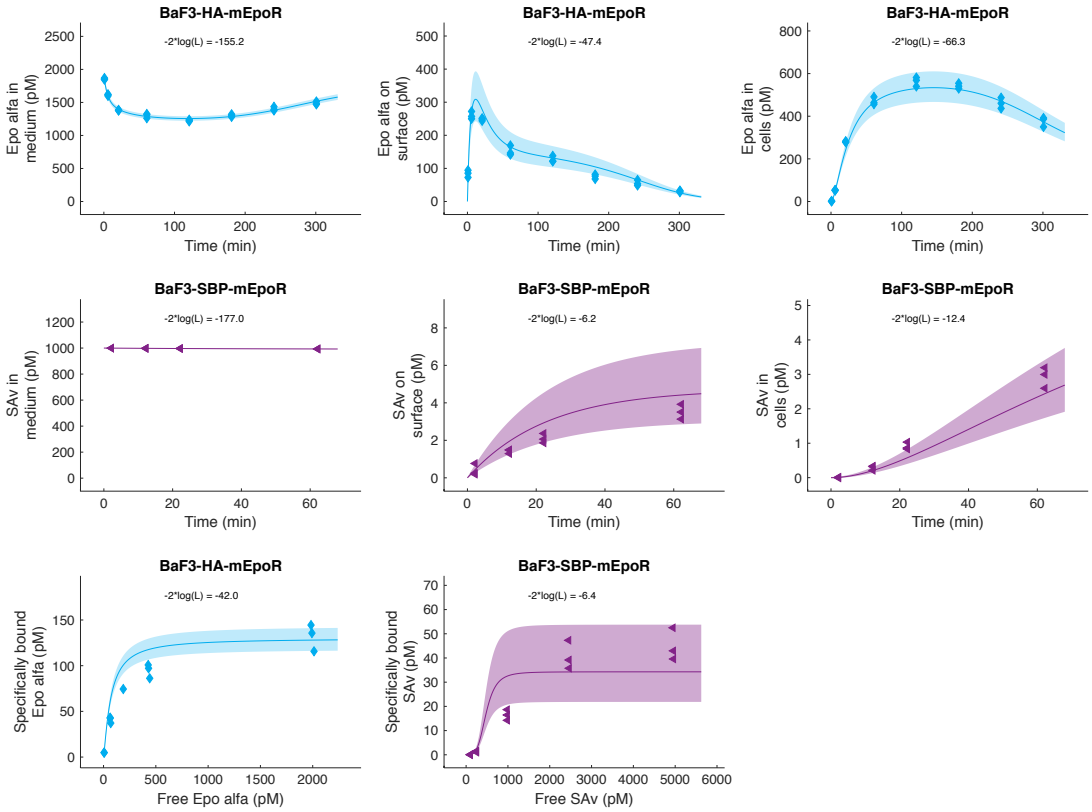

Fig. 5.3.: Model simulations and data for the revised model and data by Becker *et al.* [12]. Lines and symbols in light blue represent BaF3-HA-mEpoR cells stimulated with Epo alfa, lines and symbols in purple depict BaF3-SBP-mEpoR cells stimulated with SAv. Shading represents standard deviation of the data as estimated by a parametric error model. Additional replicates are shown in Figure B.1 in the appendix.

While the model curves for Epo alfa match the data well, for the SAv data, deviations can be observed. These deviations most probably originate from the fact, that internal model parameters differ between the differently tagged cell lines (BaF3-HA-mEpoR and BaF3-SBP-mEpoR) that have not been taken into account in the auxiliary model for SAv dynamics. Nevertheless, a good overall agreement of model and data is obtained.

Depletion of four different ESAs in the medium of BaF3 cells

Based on the data by Becker *et al.* [12], the number of ESA binding sites and the affinities of Epo alfa to the murine EpoR as well as the kinetic parameters of the core model could be determined. The kinetic parameters of the core model accounting for ESA-EpoR internalization, ESA degradation and receptor turn-over are assumed to be independent of the ligand used for stimulation. Therefore, the calibrated model can be utilized to characterize the binding properties of novel ESAs.

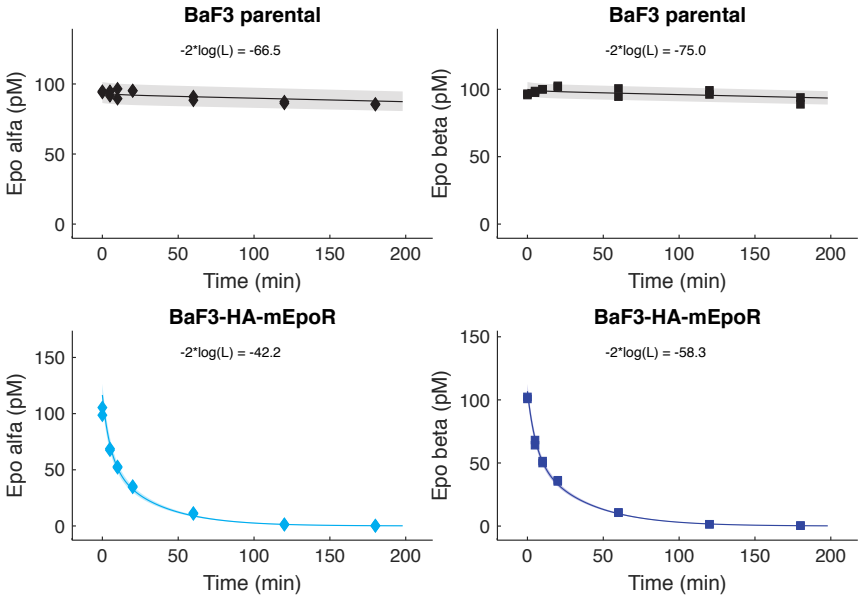

Fig. 5.4.: Depletion of ESA in the supernatant of BaF3 cells. Colors indicate ESA (Epo alfa, light blue; Epo beta, dark blue). Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model. Epo alfa and beta show similar depletion dynamics in BaF3-HA-mEpoR cells. In parental BaF3 cells, not expressing the EpoR, no depletion can be observed. Additional replicates are shown in Figure B.2 in the appendix.

A set of four ESAs was chosen to be investigated more closely: Epo alfa, Epo beta, NESP and C.E.R.A.. The two recombinant Epo variants alfa and beta have the same amino acid sequence as the endogenous version and show a very similar binding affinity [54, 73]. More recently, two genetically modified variants, called *novel erythropoiesis stimulating agent* (NESP) and *continuous erythropoiesis receptor activator* (C.E.R.A.), have been engineered that offer an easier dosing schedule due to a drastically increased half life and potency *in vivo* [120].

First, the depletion of Epo alfa and beta in the BaF3 cell line carrying the murine EpoR (BaF3-HA-mEpoR) and in parental BaF3 cells is analyzed, as shown in Figure 5.4. While the stimulated parental BaF3 cells show almost no change in ESA concentration over time (upper panels), the cells expressing EpoR rapidly deplete the ligand from the medium (lower panels). This indicates that the main proportion of ESA depletion is due to internalization of receptor-bound ligand.

The shape of the decay depends very much on the number of binding sites, represented by the model parameter $\mathrm{ESA}_{\mathrm{bind}}$, and on the receptor's binding affinity for a certain ligand, represented by the forward rate constant k_{on} and the backward rate constant k_{off}. For the data shown in Figure 5.4, both ESAs have similar affinities whereas the number

Tab. 5.2.: ESA affinities for mEpoR as determined by the ESA-EpoR model. Confidence intervals are based on the profile likelihood method. The corresponding likelihood profiles are shown in Figure B.6 in the appendix.

ESA	Parameter	Unit	Affinity
Epo alfa	k_{on}	$\mathrm{pM}^{-1}\,\mathrm{min}^{-1}$	$(2.215 \pm 0.173) \times 10^{-4}$
	k_{off}	min^{-1}	$(4.449 \pm 0.760) \times 10^{-2}$
	K_{D}	pM	$(2.009 \pm 0.172) \times 10^{2}$
Epo beta	k_{on}	$\mathrm{pM}^{-1}\,\mathrm{min}^{-1}$	$(2.000 \pm 0.310) \times 10^{-4}$
	k_{off}	min^{-1}	$(4.604 \pm 1.435) \times 10^{-2}$
	K_{D}	pM	$(2.302 \pm 0.313) \times 10^{2}$

of binding sites differs substantially between the parental and the transfected BaF3 cell line.

Using the ESA-EpoR model, these differences can be quantified computationally. For BaF3-HA-mEpoR cells, we find an estimated number of ESA binding sites of 7827 ± 377 which is perfectly in line with the result of 7769 ± 361 determined by a conventional saturation binding assay [12]. For the control, the parental BaF3 cell line, only 29 ± 9 binding sites per cell are estimated by the model, confirming that no or only few receptors are present. The affinity values of Epo alfa and beta with respect to the murine Epo receptor (mEpoR) are denoted in Table 5.2. Both rate constants, k_{on} and k_{off}, deviate only by about $10\,\%$ between Epo alfa and beta leading to very similar depletion dynamics.

Next, the affinities of all four ligands with respect to the human Epo receptor are analyzed. The data and model fits are depicted in Figure 5.5. Again, the upper panels show the control experiments with parental BaF3 cells and the lower panels indicate the depletion curves of the respective ESA in BaF3-HA-hEpoR cells.

The estimated affinities are given in Table 5.3. Comparing the values for mEpoR, an about 4-fold larger k_{on} rate for the human receptor is found for both ligands. For the k_{off}-rate, the human affinity is by a factor of two smaller than in the case of the murine receptor. Looking at the ratio of both rates, which corresponds to the K_{D}-value, a 10-fold smaller value for hEpoR than for mEpoR is observed.

With these differences in mind, the two novel molecules NESP [51] and C.E.R.A. [121] are added to the analysis. Both molecules were introduced with the claim of a greatly prolonged half-life and potency in *in vivo*. For a short overview of the biochemical properties of different ESAs, I refer to the review by Macdougall [120].

As indicated in Table 5.3, the K_{D}-values of NESP and C.E.R.A. are indeed around 30 and 50-fold higher than for Epo alpha and beta, making the depletion of the ligands from medium much slower and therefore prolonging the half-life. This 50-fold difference for

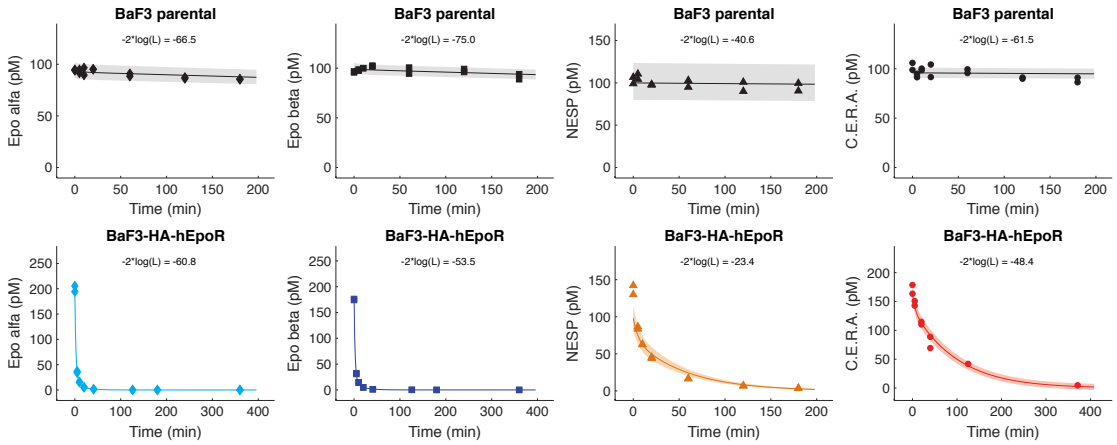

Fig. 5.5.: Depletion of ESA in the supernatant of BaF3 cells. Colors indicate ESA (Epo alfa, light blue; Epo beta, dark blue; NESP, orange; C.E.R.A., red). Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model. Additional replicates are shown in Figure B.3 in the appendix.

Tab. 5.3.: ESA affinities for hEpoR as determined by the ESA-EpoR model. Confidence intervals are based on the profile likelihood method. The corresponding likelihood profiles are shown in Figure B.6 in the appendix.

ESA	Parameter	Unit	Affinity
Epo alfa	k_{on}	$\mathrm{pM}^{-1}\,\mathrm{min}^{-1}$	$(8.421 \pm 0.536) \times 10^{-4}$
	k_{off}	min^{-1}	$(1.797 \pm 0.313) \times 10^{-2}$
	K_{D}	pM	$(2.134 \pm 0.221) \times 10^{1}$
Epo beta	k_{on}	$\mathrm{pM}^{-1}\,\mathrm{min}^{-1}$	$(9.186 \pm 0.551) \times 10^{-4}$
	k_{off}	min^{-1}	$(1.981 \pm 0.299) \times 10^{-2}$
	K_{D}	pM	$(2.157 \pm 0.185) \times 10^{1}$
NESP	k_{on}	$\mathrm{pM}^{-1}\,\mathrm{min}^{-1}$	$(1.257 \pm 0.443) \times 10^{-4}$
	k_{off}	min^{-1}	$(8.362 \pm 4.639) \times 10^{-2}$
	K_{D}	pM	$(6.650 \pm 1.055) \times 10^{2}$
C.E.R.A.	k_{on}	$\mathrm{pM}^{-1}\,\mathrm{min}^{-1}$	$(8.132 \pm 1.720) \times 10^{-5}$
	k_{off}	min^{-1}	$(8.810 \pm 3.265) \times 10^{-2}$
	K_{D}	pM	$(1.083 \pm 0.142) \times 10^{3}$

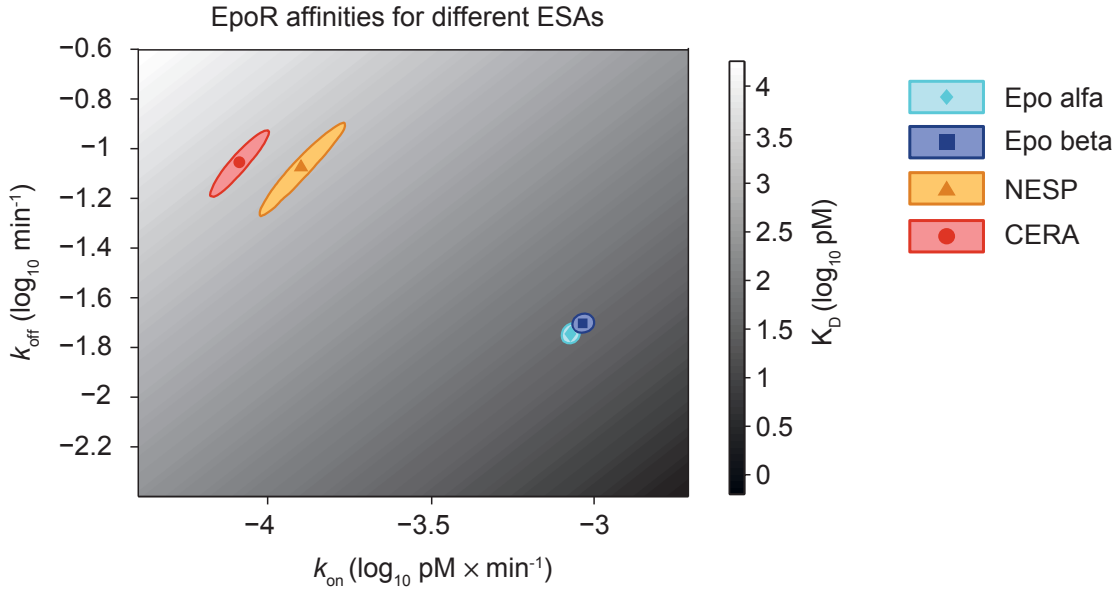

Fig. 5.6.: Two-dimensional identifiability of different ESAs. The values of the affinity parameters k_{on} and k_{off} for different ESAs are shown by symbols. Shading indicates the $68\,\%$-confidence intervals based on the profile likelihood method. The gradient in the background indicates the corresponding K_{D} value. The corresponding one-dimensional likelihood profiles are shown in Figure B.6 in the appendix.

C.E.R.A. is in line with previous results obtained in a binding assay based on a surface plasmon resonance measurement using an immobilized EpoR [93].

In order to directly compare all ligand affinities, two-dimensional likelihood profiles for the parameters k_{on} and k_{off} are calculated. In Figure 5.6, this affinity landscape is shown for all four ESAs. While the confidence regions for Epo alfa and beta overlap, NESP and C.E.R.A. show distinct affinity values with C.E.R.A. showing to lowest k_{on} value and thus the highest K_{D} value.

Utilizing the calibrated model to characterize cell lines

After having calibrated the affinities of all four ESAs to the EpoR in BaF3 cells, further depletion experiments with one of the previously characterized ESAs can be performed in other cell lines. As the internal model parameters and the affinities are identified, the only remaining unknown quantity is the cell line-specific number of binding sites, represented by the parameter $\mathrm{ESA}_{\mathrm{bind}}$. Utilizing the calibrated model, this number can be determined, which is of great use because the number of receptors on the cell surface is experimentally difficult to obtain.

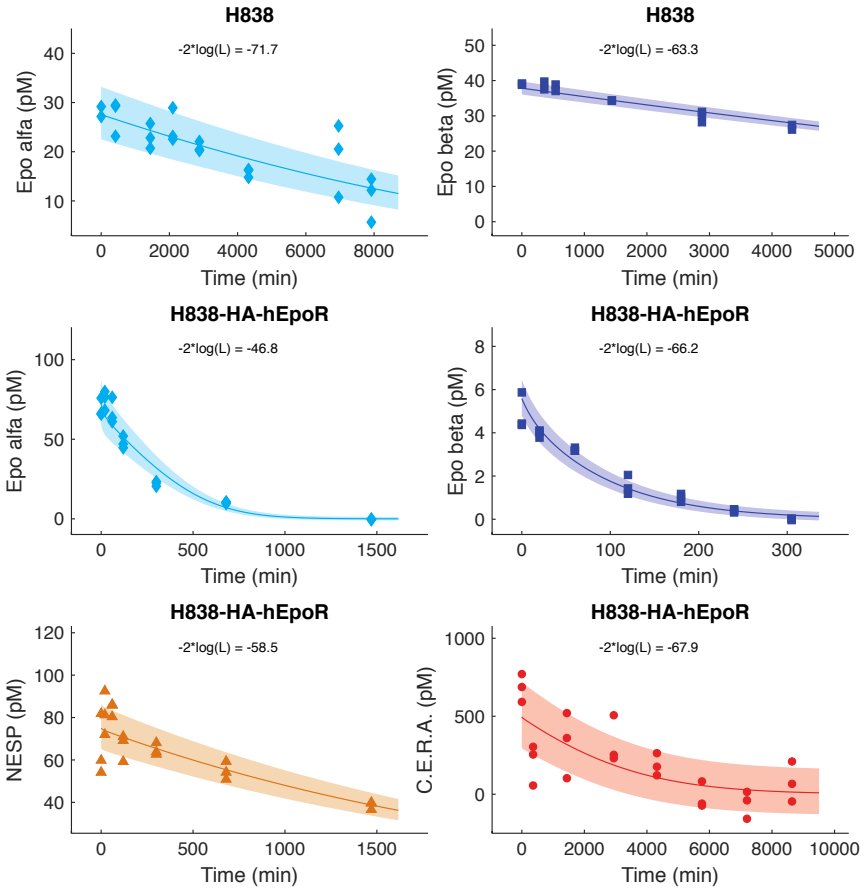

Fig. 5.7.: Depletion of ESA in the supernatant of the NSCLC cell line H838. Colors indicate ESA (Epo alfa, light blue; Epo beta, dark blue; NESP, orange; C.E.R.A., red). Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model. Additional replicates are shown in Figure B.4 in the appendix.

For the *non-small-cell lung carcinoma* (NSCLC) cell line H838 and other NSCLC cell lines, it was shown that JAK-STAT signaling can be induced upon stimulation with an ESA [49] suggesting that functional Epo receptors are present on the cell surface. *Quantitative real-time PCR* (qRT-PCR) showed that low levels of *EPOR* mRNA are detectable in H838 cells and quantitative immunoblotting revealed that the EpoR protein is phosphorylated $10\,\mathrm{min}$ after stimulation with $10\,\mathrm{U/ml}$ of Epo beta, as depicted in Figure B.8 in the appendix.

In the upper panels of Figure 5.7, experimental data and the corresponding model fits are depicted for H838 cells stimulated with Epo alfa and beta. Because of the low expression levels of EpoR, the depletion in the medium was monitored for up to $146\,\mathrm{h}$ ($8760\,\mathrm{min}$).

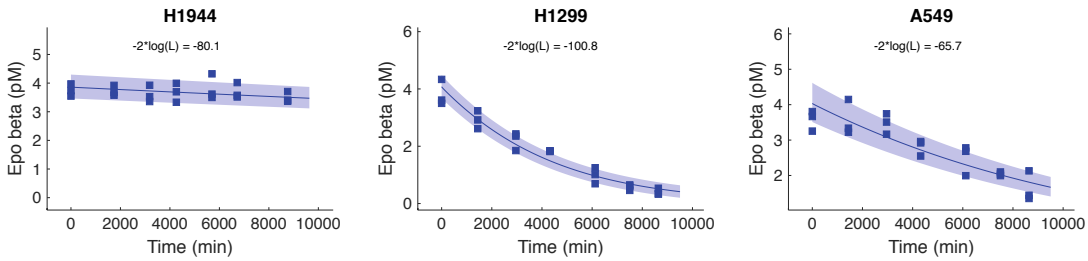

Fig. 5.8.: Depletion of Epo beta in the supernatant of NSCLC cells. Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model. Additional replicates are shown in Figure B.4 and Figure B.5 in the appendix.

The estimated number of binding sites in these cells is 82 ± 6 per cell and thus almost 100-fold lower than in BaF3-HA-EpoR cells but sill three times higher than in parental BaF3 cells. Epo alfa and beta have, as previously discussed, a higher affinity for the hEpoR than the novel ligands NESP and C.E.R.A.. For this reason, the depletion of NESP and C.E.R.A. in H838 cells is even slower than with Epo alfa and beta. By adding the transfected cell line H838-HA-hEpoR, that over expresses the human EpoR, to the panel, also the interaction with the two novel ligands can be investigated, as shown in the lower panels of Figure 5.7. For these cells, 6330 ± 138 ESA binding sites per cell are found which is only slightly lower than for BaF3-HA-EpoR cells and almost 80-fold higher than in untransfected H838 cells.

The panel of NSCLC cell lines is extended to H1944, H1299 and A549 cells. Previous analyses using qRT-PCR showed that while H1299 cells have a non-zero gene expression of *EPOR* mRNA, H1944 and A549 are below detection limit as shown in Figure B.8 in the appendix. In quantitative immunoblotting, EpoR and pEpoR protein levels are detectable in H1299 cells and small amounts in A549 cells. In Figure 5.8 the depletion of Epo beta and the corresponding model fits are shown for these three cell lines.

For the EpoR abundance in H1944 and H1299 cells, the model predicts 7 ± 3 and 98 ± 15 binding sites, respectively, confirming the qRT-PCR data and immunoblotting data. For the A549 cell line, 46 ± 2 binding sites are found, reflecting the HA-EpoR protein levels in the immunoblot.

As a last step, the erythroid progenitor cells in the *colony-forming unit* stage (CFU-E) are analyzed (see also fig:erythropoiesis). As mentioned in the introduction of this section, these cells can differentiate into erythroblasts upon Epo stimulation and therefore need to express substantial amounts of EpoR on the cell surface. The model fit of the depletion data, as shown in Figure 5.9, provides 433 ± 23 binding sites. This is consistent with previous experiments that reported 300 to 1100 binding sites per CFU-E cell [21]. As a

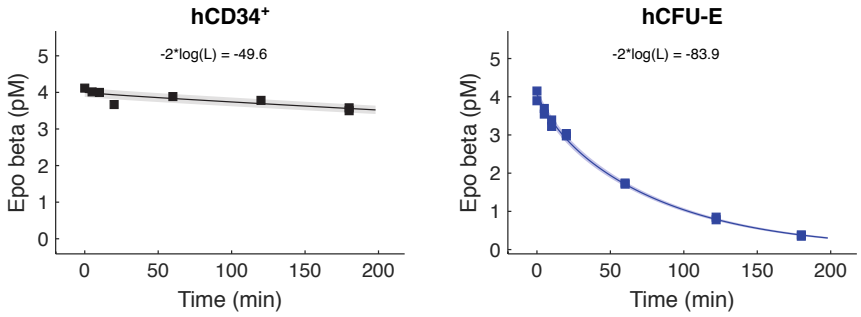

Fig. 5.9.: Depletion of Epo beta in the supernatant of blood progenitor cells. Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model. Additional replicates are shown in Figure B.5 in the appendix.

control, also the human progenitor cells hCD34$^+$, that lack the EpoR, are analyzed. As expected, the depletion in the supernatant is very slow and the corresponding number of binding sites is only 21 ± 15.

In summary, the computational model of ESA-EpoR interaction can be utilized to reliably detect and quantify the number of ESA binding sites over three orders of magnitude on various cell lines and types. Therefore, the model provides a helpful tool for characterizing cell lines and novel ESAs. A patent describing the method and its potential applications was filed in 2016 [168].

5.2.6 Linking the receptor model to JAK-STAT signaling

In the following sections, a short overview of the continuation of the project is presented. As stated in the introduction, the simulations and analyses shown here were not carried out by the author but were included in order to show the bigger picture of the approach.

The number of ESA binding sites on a specific cell type determines the capability of Epo-induced signaling. One of the main pathways activated upon Epo stimulation is the JAK2-STAT5 pathway that controls cell fate decisions. After having characterized the binding properties of four different ESAs and the EpoR abundance on the surface of various cell lines, we want to link ligand-induced receptor activation to downstream signaling.

In the last two decades, several computational models of JAK-STAT signaling have been published [189, 8, 131]. Only recently, Merkle *et al.* [131] provided a model of Epo-induced JAK2-STAT5 signaling that is able to capture the downstream dynamics in multiple cell types including the NSCLC cell line H838 and erythroid progenitor cells in the CFU-E stage. While the model by Merkle *et al.* [131] offers a great level of detail, here we limit ourselves to a very reduced JAK-STAT model that is still able to link a given Epo stimulus to a transcriptional read-out. This extended ESA-EpoR-STAT5 model is depicted in Figure 5.10a. It consists of the ESA-EpoR core model as described in (5.1)–(5.6) and two additional equations for STAT5 and phosphorylated STAT5 (pSTAT5) as defined in (5.22)–(5.23):

$$\frac{\mathrm{d}[\mathrm{STAT5}]}{\mathrm{d}t} = -[\text{ESA-EpoR}]^2 \cdot [\mathrm{STAT5}] \cdot k_{\mathrm{STAT5,act}} + [\mathrm{pSTAT5}] \cdot k_{\mathrm{STAT5,deact}} \tag{5.22}$$

$$\frac{\mathrm{d}[\mathrm{pSTAT5}]}{\mathrm{d}t} = +[\text{ESA-EpoR}]^2 \cdot [\mathrm{STAT5}] \cdot k_{\mathrm{STAT5,act}} - [\mathrm{pSTAT5}] \cdot k_{\mathrm{STAT5,deact}} \tag{5.23}$$

Based on time course data of Epo beta-induced STAT5 phosphorylation in CFUE-E and H838 cells, as shown in Figure 5.10b, the two additional parameters $k_{\mathrm{STAT5,act}}$ and $k_{\mathrm{STAT5,deact}}$ can be calibrated. By measuring the pSTAT5 level in CFU-E cells in response to increasing doses of Epo beta and C.E.R.A., the equipotent dose, which is the dose at which pSTAT5 is half-maximal, can be determined for both ESAs (Epo beta $1331\,\mathrm{pM}$, C.E.R.A. $8841\,\mathrm{pM}$), as depicted in Figure 5.10c.

When simulating the model for these equipotent doses, the model predicts a similar area under curve (AUC) for both ESAs in the case of CFU-E cells, as shown in Figure 5.10d,

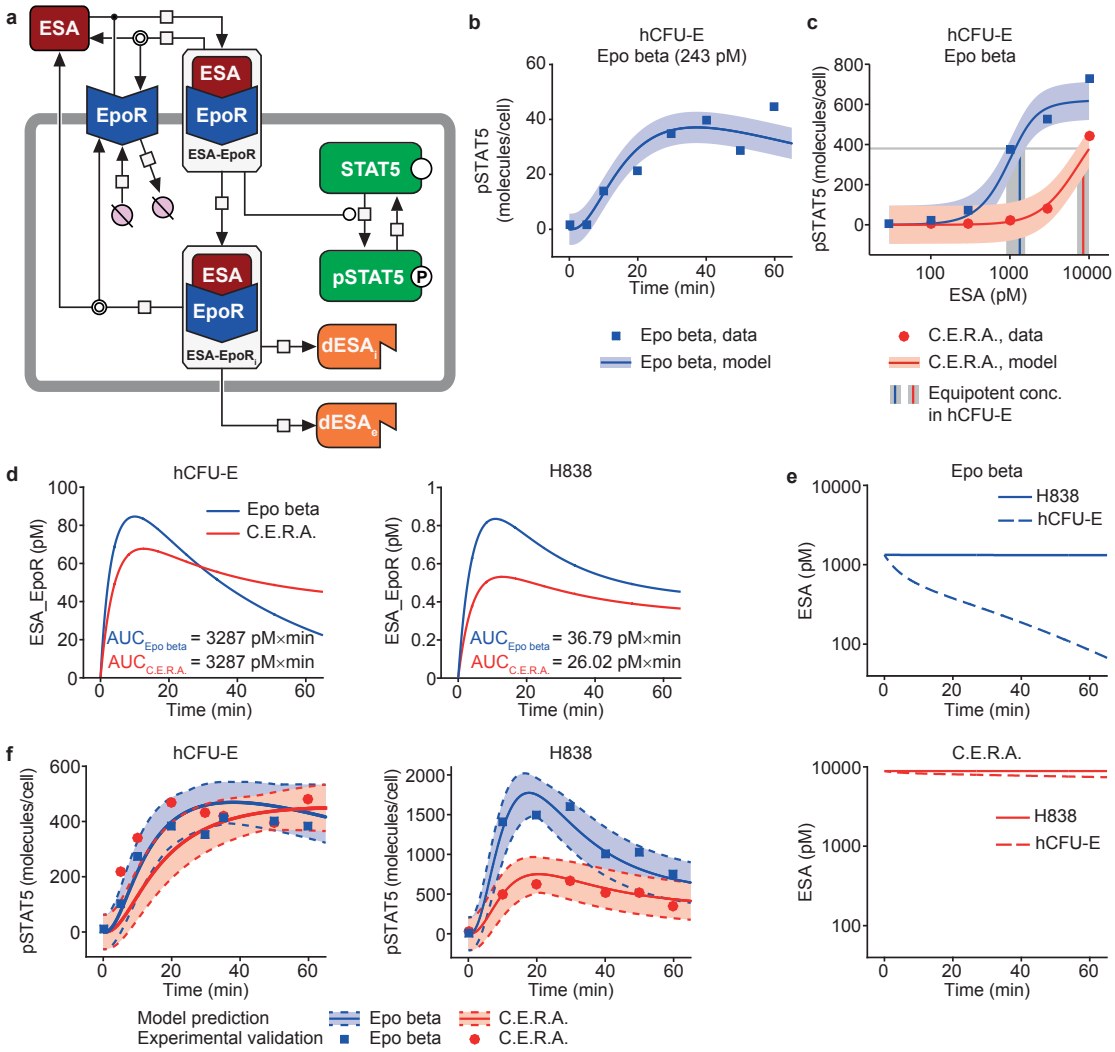

Fig. 5.10.: Impact of differences in the abundance of EpoR and of ESA binding properties on the dynamics of EpoR activation. **(a)** Graphical representation of the extended ESA-EpoR-STAT5 model. **(b)** Epo beta-induced dynamics of STAT5 phosphorylation was determined by mass spectrometry for hCFU-E. **(c)** Mass spectrometric quantification of Epo beta and C.E.R.A. dose-dependent induction of STAT5 phosphorylation in hCFU-E at 10 min. Equipotent concentrations are indicated with vertical lines. **(d)** Model-predicted time courses and area-under-curve (AUC) of EpoR occupancy for Epo beta and C.E.R.A. in hCFU-E and H838 cells. **(e)** Model-predicted depletion of Epo beta and C.E.R.A. by hCFU-E cells and H838. **(f)** Dynamics of STAT5 phosphorylation induced by equipotent concentrations of Epo beta and C.E.R.A.. Model-predicted time courses and absolute quantification of phospho-STAT5 molecules per cell in hCFU-E and H838 cells by mass spectrometry are shown. Solid lines represent model trajectories and shading reflects the standard deviation of the data, shadings delimited with dashed lines indicate validation profile likelihood-based confidence band of model predictions. Figure adapted from Rodriguez-Gonzalez *et al.* [170].

which corresponds to a similar pathway activation. For the H838 cells, however, the same dose lead to a different AUC for C.E.R.A.. This is due to the lower number of ESA binding sites on H838 cells and the decreased affinity of C.E.R.A. as compared to Epo beta. This can also be seen when simulating the depletion of Epo beta and C.E.R.A. in CFU-E and H838 cells, as shown in Figure 5.10e. While for Epo beta, a strong depletion is observed in CFU-E cells, H838 cells do not visibly deplete the ligand from the medium. By contrast, for C.E.R.A., only a minor depletions occurs in CFU-E cells and in H838 cells the concentration remains constant.

Model predictions for the number of phosphorylated STAT5 molecules in hCFU-E and H838 cells during $60\,\mathrm{min}$ after stimulation with Epo beta and C.E.R.A. are indicated by the dashed areas in Figure 5.10f. Here, similar pSTAT levels in CFU-E cells in response to Epo beta and C.E.R.A. are observed. By contrast, in H838 cells, C.E.R.A. induces a weaker activation than Epo beta. These predictions were successfully validated using mass-spectrometry, as represented by the symbols in Figure 5.10f.

5.2.7 Pharmacokinetics and -dynamics (PKPD)

In the previous parts, we could show that NSCLC cells can have signaling competent ESA binding sites and that C.E.R.A. is less active in H838 cells than the equipotent dose of Epo beta. In order to transfer these results to the patient level, another extended version of the ESA-EpoR core model was built. This multi-scale model comprises the dynamics of ESA in the interstitium (pharmacokinetics; PK) and also models hemoglobin degradation as a function of the concentration of the ESA-EpoR complex (pharmacodynamics; PD). The combined ESA-EpoR-PKPD model can be used to determine patient specific parameters based on longitudinal data from clinical trials.

The model structure is depicted in Figure 5.11a. The corresponding equation system consists of the core ESA-EpoR model converted to the hours time scale (5.26)–(5.31) as well as two additional equations for the hemoglobin ($[\mathrm{Hb}]$) level in the blood (5.24) and the ESA concentration in the interstitium ($[\mathrm{ESA_{SC}}]$) (5.25).

$$\frac{\mathrm{d}[\mathrm{Hb}]}{\mathrm{d}t} = +[\text{ESA-EpoR}] \cdot k_{\text{hb\_pro}} - [\mathrm{Hb}] \cdot k_{\text{hb\_deg}} \cdot k_{\text{hb\_deg\_PatientID}} \tag{5.24}$$

$$\frac{\mathrm{d}[\mathrm{ESA_{SC}}]}{\mathrm{d}t} = +[\text{Injection}] - [\mathrm{ESA_{SC}}] \cdot k_{\text{sc\_out}} - \frac{[\mathrm{ESA_{SC}}] \cdot k_{\text{sc\_clear}}}{[\mathrm{ESA_{SC}}] + k_{\text{sc\_clear\_sat}}} \tag{5.25}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{ESA}]}{\mathrm{d}t} = &+[\mathrm{ESA_{SC}}] \cdot k_{\text{sc\_out}} - [\mathrm{ESA}] \cdot k_{\text{clear}} - 60 \cdot [\mathrm{ESA}] \cdot [\mathrm{EpoR}] \cdot k_{\text{on}} \\ &+ 60 \cdot [\text{ESA-EpoR}] \cdot k_{\text{off}} + 60 \cdot [\text{ESA-EpoR}_\text{i}] \cdot k_{\text{ex}}\end{aligned} \tag{5.26}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{EpoR}]}{\mathrm{d}t} = &-60 \cdot [\mathrm{ESA}] \cdot [\mathrm{EpoR}] \cdot k_{\text{on}} + 60 \cdot [\text{ESA-EpoR}] \cdot k_{\text{off}} \\ &+ 60 \cdot \mathrm{ESA_{bind}} \cdot k_{\text{t}} - 60 \cdot [\mathrm{EpoR}] \cdot k_{\text{t}} + 60 \cdot [\text{ESA-EpoR}_\text{i}] \cdot k_{\text{ex}}\end{aligned} \tag{5.27}$$

$$\begin{aligned}\frac{\mathrm{d}[\text{ESA-EpoR}]}{\mathrm{d}t} = &+60 \cdot [\mathrm{ESA}] \cdot [\mathrm{EpoR}] \cdot k_{\text{on}} - 60 \cdot [\text{ESA-EpoR}] \cdot k_{\text{off}} \\ &- 60 \cdot [\text{ESA-EpoR}] \cdot k_{\text{e}}\end{aligned} \tag{5.28}$$

$$\begin{aligned}\frac{\mathrm{d}[\text{ESA-EpoR}_\text{i}]}{\mathrm{d}t} = &+60 \cdot [\text{ESA-EpoR}] \cdot k_{\text{e}} - 60 \cdot [\text{ESA-EpoR}_\text{i}] \cdot k_{\text{ex}} \\ &- 60 \cdot [\text{ESA-EpoR}_\text{i}] \cdot k_{\text{di}} - 60 \cdot [\text{ESA-EpoR}_\text{i}] \cdot k_{\text{de}}\end{aligned} \tag{5.29}$$

$$\frac{\mathrm{d}[\mathrm{dESA_i}]}{\mathrm{d}t} = +60 \cdot [\text{ESA-EpoR}_\text{i}] \cdot k_{\text{di}} \tag{5.30}$$

$$\frac{\mathrm{d}[\mathrm{dESA_e}]}{\mathrm{d}t} = +60 \cdot [\text{ESA-EpoR}_\text{i}] \cdot k_{\text{de}} \tag{5.31}$$

The model is calibrated to longitudinal data of several clinical trials: two phase II trials conducted with C.E.R.A. in NSCLC patients suffering from normocytic anemia (NCT00072059 [82], NCT00327535 [27]) and two phase I trials in healthy subjects [97, 166]. In all trials, the C.E.R.A. concentration and the Hb level were monitored for up to 12 weeks. While the parameters of the core model were defined by the previous calibration based on *in vitro* data, the additional parameters needed to be estimated either for the whole patient population or in a patient-specific manner. A model comparison using the likelihood ratio test revealed that only the number of ESA binding sites ($\mathrm{ESA_{bind}}$) and the Hb degradation rate $k_{\text{hb\_deg\_PatientID}}$ need to be patient-specific while all other parameters were compatible over all subjects from the four clinical trials.

In Figure 5.11 the data and model fits are shown for two representative patients, one from NSCLC (panel b) and one from the healthy group (panel c). In Figure 5.11d, the histograms for the whole patient population is given. In general, an increase of the Hb degradation rate was observed in the NSCLC patient population while the number of ESA binding sites decreased compared to the healthy subjects.

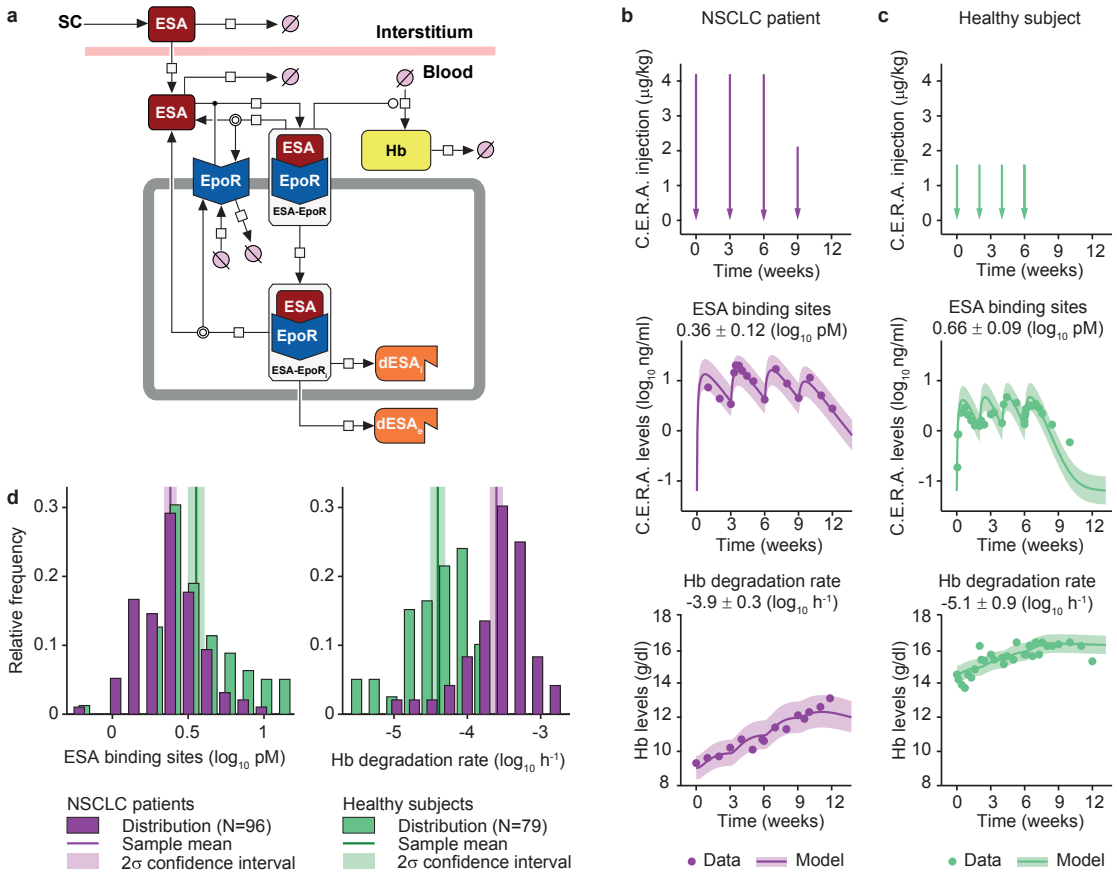

Fig. 5.11.: Representation of clinical data by a multi-scale ESA-EpoR-PK/PD model and patient-specific determination of ESA binding sites and hemoglobin degradation rates. (a) Graphical representation of the multi-scale ESA-EpoR-PK/PD model. Hb: hemoglobin, sc: subcutaneous. **(b)** Timing of C.E.R.A. doses (arrows), pharmacokinetics of C.E.R.A. and pharmacodynamics of Hb levels for a NSCLC patient (purple) (ID:2101, clinical trial NCT00072059). **(c)** same as b) but with healthy subject (green) (ID:25, CSR WP16422 clinical trial). Experimental measurements are indicated by dots. Solid lines represent the model trajectories and shading the standard deviation of the data. The model-estimated ESA binding sites and the Hb degradation rate ± confidence intervals for each subject are given. **(d)** Model estimated distributions of ESA binding sites and of the Hb degradation rates in NSCLC patients (NCT00072059 clinical trial) in purple (N=96) and in healthy subjects (CSR WP16422 and CSR BP18035 clinical trials) in green (N=79). For both parameters, Wilcoxon rank analysis revealed $p < 1 \times 10^{-5}$. Figure adapted from Rodriguez-Gonzalez *et al.* [170].

Model-based risk prediction

Although patients frequently suffer from anemia associated with chemotherapy, treating them with ESAs has been associated with an increase in the risk of mortality [192]. Here, the patient outcome of two clinical trails in NSCLC patients is correlated with ESA treatments using the patient-specific model parameters for Hb degradation and the number of ESA binding sites. Applying Youden's J statistic, the prognostic value of both parameters and the given dose of ESA can be defined. The model was able to classify the patients into high or low risk of fatal outcome and showed a good correlation between patients in the high risk group and the deceased patients, as shown in Figure 5.11a. The prediction was validated using only PD values and the ESA doses in an independent group of patients. The mathematical model was capable to correctly assign the relative majority of the deceased patients to the high risk group of fatal outcome, as depicted in Figure 5.11b.

Predicting an optimal treatment strategy for individual patients

The response of individual patients to an ESA can be predicted by the model after the first doses have been administered. Therefore, the multi-scale ESA-EpoR-PKPD model can also be utilized to define optimal dosing schedules only a few weeks after the initial treatment. The process is depicted in Figure 5.11c–f. When treating patients according to the guidelines, as shown in panel c) and d), some patients might be over or under dosed. In contrast, the model-based optimized treatment can correct the dosing, as shown in panel e). Only a minority of the patients do not respond strong enough, therefore requiring a more densely dosing schedule, as indicated in panel f).

5.2.8 Discussion

In sum, based on the Epo-EpoR model by Becker *et al.* [12] a mechanistic multi-scale model could be established that is able to transfer the knowledge about ligand and cell characteristics obtained from *in vitro* data to the dosing schedule for patients in clinics.

The first outcome of the project is, that the revised ESA-EpoR model is able to describe the dynamics and thereby determine the ligand-receptor affinities of four different ESAs in BaF3 cells stably expressing the murine or human EpoR. This calibrated model can be utilized as a tool to characterize other cell lines and ligands. This can be of great value for drug development as the standard approaches for characterizing ESAs like BIAcore suffer from the artificial experimental conditions (e.g. the receptors are mounted on a metal

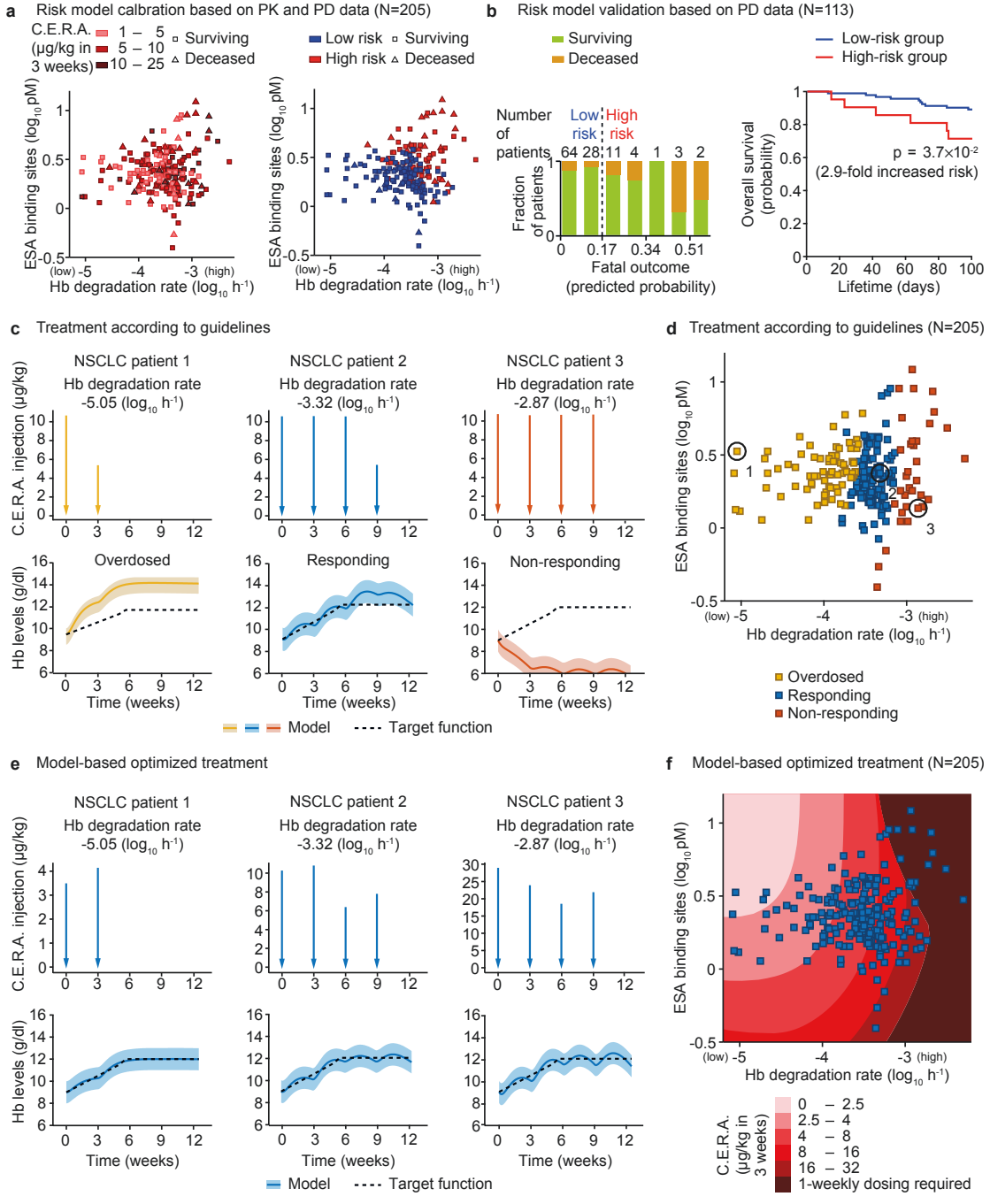
a
Risk model calbration based on PK and PD data (N=205)
C.E.R.A. (µg/kg in 3 weeks)
1 – 5
5 – 10
10 – 25
Surviving
Deceased
Low risk
High risk
ESA binding sites (log10 pM)
Hb degradation rate (log10 h-1)
(low)
(high)
b
Risk model validation based on PD data (N=113)
Surviving
Deceased
Number of patients
Low risk
High risk
64 28 11 4 1 3 2
Fraction of patients
Fatal outcome (predicted probability)
Low-risk group
High-risk group
Overall survival (probability)
p = 3.7×10-2
(2.9-fold increased risk)
Lifetime (days)
c
Treatment according to guidelines
NSCLC patient 1
NSCLC patient 2
NSCLC patient 3
Hb degradation rate -5.05 (log10 h-1)
Hb degradation rate -3.32 (log10 h-1)
Hb degradation rate -2.87 (log10 h-1)
C.E.R.A. injection (µg/kg)
Overdosed
Responding
Non-responding
Hb levels (g/dl)
Time (weeks)
Model
Target function
d
Treatment according to guidelines (N=205)
ESA binding sites (log10 pM)
Hb degradation rate (log10 h-1)
Overdosed
Responding
Non-responding
e
Model-based optimized treatment
NSCLC patient 1
NSCLC patient 2
NSCLC patient 3
Hb degradation rate -5.05 (log10 h-1)
Hb degradation rate -3.32 (log10 h-1)
Hb degradation rate -2.87 (log10 h-1)
C.E.R.A. injection (µg/kg)
Hb levels (g/dl)
Time (weeks)
Model
Target function
f
Model-based optimized treatment (N=205)
ESA binding sites (log10 pM)
Hb degradation rate (log10 h-1)
C.E.R.A. (µg/kg in 3 weeks)
0 – 2.5
2.5 – 4
4 – 8
8 – 16
16 – 32
1-weekly dosing required

Fig. 5.11.: Model-based response optimization and risk prediction. (a) Patient stratification based on individual estimations of ESA binding sites and Hb degradation rate in both NSCLC clinical trials (NCT00072059 and NCT00327535). Patients who died during the trial are depicted as triangles, and survivors as squares. In the left panel, the color code indicates the ESA dose given in the first three weeks. In the right panel, red and blue colors indicate the classification of patients in high risk and low risk of fatal outcome. **(b)** Representation of an independent dataset of the NCT00072059 clinical trial for which only PD values were available. Left panel corresponds to patient groups based on their predicted probability of fatal outcome. Green and orange indicate the fraction of patients that survived or died. Dashed line indicates the risk threshold. Overall survival is displayed as a Kaplan-Meier plot (right panel), blue and red colors correspond to the predicted low and high risk group. The cox-proportional hazard model was used to determine a 2.9-fold increase in risk ($p < 0.05$). **(c)** C.E.R.A. treatment according to the guidelines for NESP for three patients with distinct Hb degradation rates (NCT00072059, patient ID 2303, 2723 and 2652) was simulated by the multi-scale model. Model-predicted Hb levels are displayed as solid lines and shading represents the standard deviation estimated from the clinical data. Dashed lines indicate the optimal Hb response according to the guidelines for ESAs. **(d)** Patient stratification based on the predicted outcome of the C.E.R.A. treated patients in the NCT00072059 and NCT00327535 trials. For each patient, the model-estimated ESA binding sites and the Hb degradation rates are depicted. Overdosed patients (yellow) are defined as $\mathrm{Hb} \geq 2\sigma$ above target response by the guidelines (dashed line). Responding patients (blue) are defined as Hb responses within the guidelines. Non-responding patients (red) are defined as $\mathrm{Hb} \geq 2\sigma$ below the target response given by the guidelines. The patients exemplified in (c) are indicated. **(e)** Model-based optimized ESA treatment of patient 1, 2 and 3. Model-predicted optimal C.E.R.A. doses and predicted Hb responses (solid lines) are indicated. Shaded areas indicate the standard deviation estimated from the clinical data and dashed lines correspond to the optimal Hb response based on the guidelines for ESAs. **(f)** Optimized ESA dosing and timing regimens are predicted by the multi-scale model for all patients depending on Hb degradation rate and ESA binding sites. ESA dosing and timing is indicated by shades of red. Figure adapted from Rodriguez-Gonzalez *et al.* [170].

surface) and can therefore deviate drastically from *in vivo* values [141]. The second feature of our method is to provide the number of ESA binding sites for the cell type under investigation. Experimentally, this number can be measured by quantitative flow cytometry (e.g. qFACS) which combines a specific antibody labeling with a calibration of the fluorescence using beads and has been applied for quantifying the abundance of subtypes of the IL1-receptor [198]. For our approach, we solely need to measure the depletion of a well characterized ESA in the medium using ELISA and our model can calculate the corresponding number.

In the second step, we linked the ESA-EpoR model to JAK-STAT signaling. Using this ESA-EpoR-JAK-STAT model we could calculate equipotent doses of different ESAs based on their *in silico* signaling activity. This allows us to perform simulations with ESAs like C.E.R.A. that have not been tested or approved for chemotherapy-induced anemia but provide more flexibility in the treatment due to their increased half-life.

The third step linked the cellular scale to the body level by extending the ESA-EpoR model by a PK/PD part. Based on the data of four clinical trails with C.E.R.A. in NSCLC patients and healthy subjects, we could calculate the patient-specific number of ESA binding sites and thereby derive the number of CFU-E cells available for differentiation. Also the patient-specific Hb degradation rate was estimated from the monitored Hb levels. The comparison of the two populations, healthy subjects and NSCLC patients, showed a general shift towards lower numbers of CFU-E cells and higher Hb degradation rates in NSCLC patients. With this information at hand, we can predict the increase of Hb caused by a dose of ESA administered to the patient and use the multi-scale model to optimize the dosing schedule within the bounds of the regulatory restrictions.

Also, the patient outcome from another study could be correlated with the three features number of ESA binding sites, Hb degradation rate and administered ESA dose. This allows for a computational risk prediction that can help to avoid under or overdosing of critical patients.

5.3 Deciphering the cellular composition of unknown patient samples for immunotherapy

This work was initiated during a six-month internship in 2015 at Merrimack Pharmaceuticals Inc. in Cambridge, MA, USA. The idea was inspired by a recent publication by Newman et al. *[139]. All steps have been developed in close collaboration with Andreas Raue (Merrimack). Sample preparation and flow cytometry analysis of the ovarian ascites samples were performed by Sonia Feau (Merrrimack) and Jinyan Du (Merrimack). Processing of the single-cell RNA-seq data of the ascites samples was performed by Nav Ranu (Merrimack). A publication of the project with the title*

***Schelker, M.**, Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B.* & *Raue, A. „Estimation of immune cell content in tumour tissue using single-cell RNA-seq data“.* Manuscript in preparation *2017*

is currently in preparation and a preprint is available on bioRXiv[1].

5.3.1 Introduction

In the five years between 2009 and 2014, 51 new anti-cancer drugs have been approved by the FDA [123] – among them, the first generation of immunotherapeutical drugs. These novel drugs make use of the body's inherent immune system to fight malignant tissue. The approach brings potentially less side effects and is, for the appropriate patient population, very effective. The selection of the right treatment strategy for individual patients, however, is still challenging as information on the presence of immune cells in the tumor environment can only be obtained by histology of tissue samples or based on well-studied biomarkers that correlate with the infiltration of specific cell types in the tumor environment.

In this section, I want to present a computational method for quantification of immune and stromal cell content in bulk tumor tissue samples. First, the methodology of mixture deconvolution is introduced and existing methods and their limitations are mentioned. Next, single-cell RNA sequencing (scRNA-seq) data from three different human tissue sources, i.e. melanoma, ovarian carcinoma ascites and peripheral mononuclear blood cells (PBMCs) from healthy subjects, are analyzed to identify and assign the major cell

[1] `http://biorxiv.org/content/early/2017/04/12/127001`

types contained in the samples. Using the combined data set, a benchmark setting is established that allows for the systematic evaluation of patient heterogeneity and its impact on the deconvolution result. Using this benchmark, we can asses the performance of different deconvolution approaches and configurations systematically and select the best performing methods for further analyses.

Besides the estimation of immune and stromal cell fractions, this revised deconvolution method allows for the estimation of the fraction of tumor cells and can predict the gene expression profile of the tumor cells in a patient-specific manner.

Tumor infiltrating immune cells

The human immune system consists of a complex network of highly specialized, interacting cell types. While cytotoxic cells, such as CD8+ T cells, fight malignant tissue by inducing programmed cell death, regulatory T cells suppress the immune response to prevent, for instance, autoimmune disease. Therefore, the balance of effector and suppressor cells is crucial for the correct functioning of the immune system. In cancer patients, this balance is perturbed and suppressor cells protect the tumor from being fought by the immune response [199]. The whole cycle of cancer-immunity, as proposed by Chen & Mellman [28], is depicted in Figure 5.12 where each of the numbered boxes shows a location within the body and denotes the respective stimulatory and inhibitory factors.

Studies revealed that abundance and location of certain immune cells can predict patient outcome with a high level of confidence [64]. The immune infiltration of tumors therefore represents an important prognostic factor for disease progression and can help to support treatment decisions. Recent developments in immunotherapy try to restore the cancer immunity by blocking specific *immune checkpoints* such as CTLA4 and PD-L1/PD1 [149]. However, a big challenge for immunotherapy in cancer treatment remains the identification of patient populations that respond to the respective therapy. Therefore, reliable approaches for computational mixture deconvolution represent a promising way to gather information on immune cell infiltration in the tumor tissue based on measurements that are often part of the clinical standard procedure.

Deconvolution of unknown cell mixtures

The deconvolution of unknown mixtures of cells can be tackled in several ways. Here, I want to focus on deconvolution methods that are based on the linearity assumption of the gene expression of all cells comprised in the mixture i.e. the gene expression of a bulk

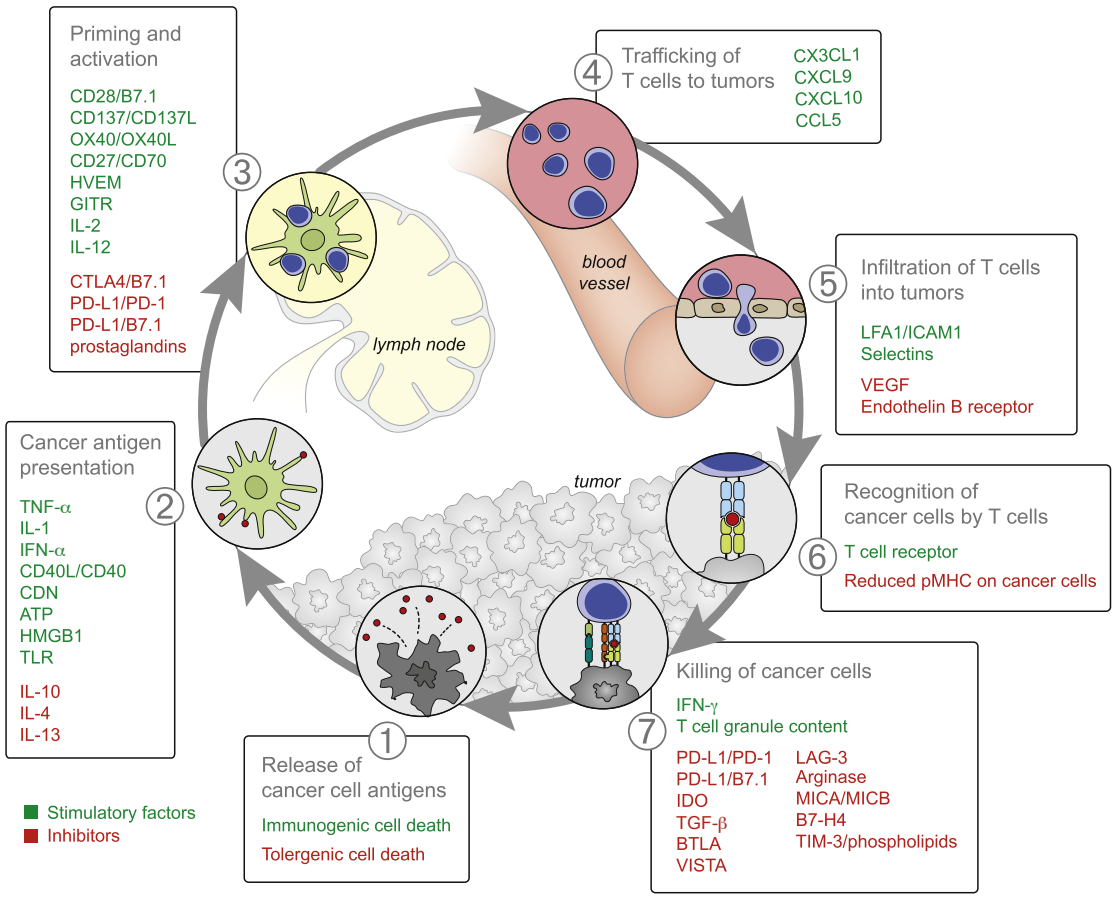

Fig. 5.12.: The cancer-immunity cycle. The body can generate immunity against cancer at multiple locations, indicated by numbers. For each step in the cycle, stimulatory and inhibitory factors are indicated. Figure taken from Chen & Mellman [28].

sample can be represented as a linear combination of the gene expression of individual cell types weighted by their proportions in the sample:

$$\vec{m}_i = B\vec{w}_i \tag{5.32}$$

where $\vec{m}_i$ is the gene expression of the mixture of the i-th patient, B is the signature matrix and is composed of one column for each cell type and $\vec{w}_i$ is the initially unknown fraction of each cell type within the mixture. Thus, if the gene expression of individual cell types is known, the system can be solved computationally for $\vec{w}_i$ and the fraction of each component can be determined.

Because the experimental data for $\vec{m}_i$ and B are obtained by measuring gene expression of single cells or cell populations, biological variability and technical measurement errors introduce noise. This noise can limit the accuracy of the deconvolution and makes the selection of genes with high discriminative power a crucial part of the deconvolution

process. Another source of uncertainty is the possible contamination of the sample with other cell types that are not contained in B. In a realistic setting, this is very relevant as most samples from patient tissue contain a variety of different cell types of which only a part is known *a priori*.

Gene expression data as a basis for deconvolution

Since microarray data of cell populations and patient samples can be obtained at a relatively low cost, the amount of available data, e.g. on NCBI's *gene expression omnibus* (GEO)[2], increases rapidly. Initiatives like the *Cancer Cell Line Encyclopedia* (CCLE) [10] and *The Cancer Genome Atlas* (TCGA) [210] have been launched and aim for an extensive collection and quantification of commonly used cancer cell lines and patient samples using various measurement and analysis techniques. More recently, high-throughput sequencing of RNA (RNA-seq) has been introduced to the community providing gene expression quantification on an absolute scale with a higher accuracy especially for low abundant RNAs (see also Section 2.1.6). These data form an ideal basis for the development and application of computational deconvolution approaches.

Existing methods for mixture deconvolution

In 2009, Abbas *et al.* purified immune cells from whole blood and measured their gene expression profiles using microarray. The authors could show that using these data the relative fractions of immune cells in independent blood samples could be accurately determined. Moreover, they applied their deconvolution approach to samples from patients suffering from the autoimmune disease *systemic lupus erythematosus* and compared the deconvolution results to healthy subjects showing that activation states of lymphocytes differ significantly between both groups. Gong *et al.* [70] used a quadratic programming approach to decipher immune cell fractions and evaluated the performance of their approach on cell line data and data from human whole-blood data from kidney transplant recipients. Qiao *et al.* [157] extended the method of Abbas *et al.* [1] from an ordinary least-squares to non-negative least-squares strategy and successfully utilized it to deconvolve whole blood samples under different micro-environmental and developmental conditions. Recently, Newman *et al.* [139] developed a method called CIBERSORT in which the authors applied *support vector regression* (SVR) to estimate cell fractions from microarray data. They also compared CIBERSORT to previously published

[2] `https://www.ncbi.nlm.nih.gov/geo/`

methods and showed that their approach outperforms existing methods when being applied to well-known *in vitro* cell mixtures. Exhaustive reviews on the topic is given by Mohammadi *et al.* [134] and Newman & Alizadeh [138].

Gentles *et al.* [67] applied the CIBERSORT method to a huge collection of publicly available data such as TCGA. Based on the deconvolution result, the authors performed a meta-analysis over 39 malignancies and identified complex associations between 22 distinct leukocyte subsets and overall survival of cancer patients.

The above-mentioned methods have been specifically developed for gene expression profiles of cell mixtures or samples and purified cell types measured by microarray. Only few attempts have been made to make use of data obtained by more recent methods like RNA-seq. Gong & Szustakowski [69] presented an algorithm for using RNA-seq data for deconvolution of immune cell fractions in patient samples. However, the proof-of-principle experiments are kept artificially small (mixture of only three known cell types and a small unknown cell fraction) and are thus only of limited value. Mehnert *et al.* [130] applied the CIBERSORT method to RNA-seq data of POLE-mutant cancers in TCGA. Based on the deconvolution result, the authors concluded that patients with POLE-mutations are good candidates for immune checkpoint inhibitor therapy. However, they did not show a proof-of-principle experiment to validate the correct operation and accuracy of CIBERSORT being utilized for RNA-seq data.

Taken together, the potential of having absolute expression values from RNA-seq measurements rather than relative data from microarray has not been fully exploited yet. Therefore, a systematic evaluation and benchmarking on realistic data is required for enabling the use of RNA-seq data for computational deconvolution.

5.3.2 Single-cell RNA-seq data from multiple tissues and patients

In the following, I want to present a proof-of-principle study for using single-cell RNA-seq (scRNA-seq) data to deconvolve patient samples based on the CIBERSORT method. For scRNA-seq measurements, the tissue samples are experimentally decomposed into single cells (see also Section 2.1.6). These cells can be classified into cell types based on their expression profiles. Thus, the ground truth of the mixture is known and can be used to assess the accuracy of the deconvolution approach.

Gathering single-cell data from multiple human tissue sources

In order to make our benchmark study robust against tissue- or patient-specific gene expression patterns, we collected data from multiple human tissues sources obtained from patient samples. We included a recently published data set of melanoma single-cell sequencing data by Tirosh *et al.* [190] which provides the gene expression levels for 4645 single cells obtained from 19 melanoma patient samples. As a second indication, we incorporated data from ascites samples of four ovarian carcinoma patients. This dataset consists of a total of 3114 individual cells sequenced by scRNA-seq and the composition of each sample was also quantified by qFACS (see also Section 2.1.2). As most of the previous deconvolution studies have been performed on immune cell data from whole blood, we also included single-cell data of PBMCs from four healthy subjects provided by 10x Genomics[3] ("4k/8k PBMCs from a Healthy Donor", "Frozen PBMCs (Donor A/B/C)" [217]). As the merged PBMC dataset comprises a total of $\sim 33\,\mathrm{k}$ individual cells, we randomly picked 1000 cells for each of the four donors to reduce the data to a comparable extent.

Overall, the combined single-cell sequencing dataset contains gene expression data for more than $17\,\mathrm{k}$ genes of $\sim 11\,\mathrm{k}$ individual cells spanning three human tissue sources.

Making the samples comparable

Gene expression values were used on the *transcripts per million* (TPM) scale as provided by current quantification methods [116, 19, 150] and were transformed to the commonly used non-negative $\log_2$-scale by

$$y = \log_2(\mathrm{TPM} + 1). \tag{5.33}$$

To ensure cross-sample comparability, all three datasets were normalized to the average expression of 3559 housekeeping genes [52] by

$$\tilde{y}_i = y_i \cdot \frac{\overline{\mathrm{HK}}}{\mathrm{HK}_i}, \tag{5.34}$$

where y_i represents the gene expression profile of the i-th sample, HK_i denotes the average gene expression over all housekeeping genes of the i-th sample and $\overline{\mathrm{HK}}$ is the average expression over all housekeeping genes and samples. Other normalization

[3] `https://support.10xgenomics.com/single-cell/datasets`

methods like upper quartile or median normalization could not be applied to scRNA-seq data as the single-cell measurements contain too many genes with zero expression leading to a zero-upper quartile and median for several samples. Gene symbols of the melanoma dataset needed to be corrected to account for automatic conversion into dates by Microsoft Excel [220].

5.3.3 Classifying individual cells based on marker gene expression and similarity

Traditional methods for classifying cells into cell types are based on cell sorting through flow cytometry (see Section 2.1.2). Therefore, specific cell surface molecules, so-called *cluster of differentiation* (CD) molecules, have been defined [221] and their expression, whether positive (e.g. CD45+) or negative, defines the differentiation stage of each cell.

For scRNA-seq data, however, the expression of the corresponding genes can be very sparse and the classification solely based on a few CD genes seems to be unreliable. Therefore, we suggest to use a combination of marker gene expression, dimensionality reduction and machine learning-based classification in order to reliably assign a cell type to each individual cell.

Dimensionality reduction and clustering

To reduce the very high-dimensional space provided by scRNA-seq data, the t-SNE algorithm (see Section 2.2.2) was applied resulting in a two dimensional representation of the data as shown in Figure 5.13a. The t-SNE mapping arranges the cells based on their similarity in gene expression. We performed the analysis on a subset of 1015 genes, that have been found to be differentially expressed in various immune cell types (see also Section 5.3.5). On this two-dimensional map, a DBSCAN clustering (see Section 2.2.3) was performed in order to assign a cluster index to each distinct group of similar cells, as indicated by colors and numbers in Figure 5.13a. The parameters were set to `MinPts=25` and `Eps=1.5`.

Initial cell type assignment based on marker gene expression

Next, we analyzed the gene expression of a total of 45 marker genes and merged the expression to an overall cell type score according to three logical operators: (1) AND

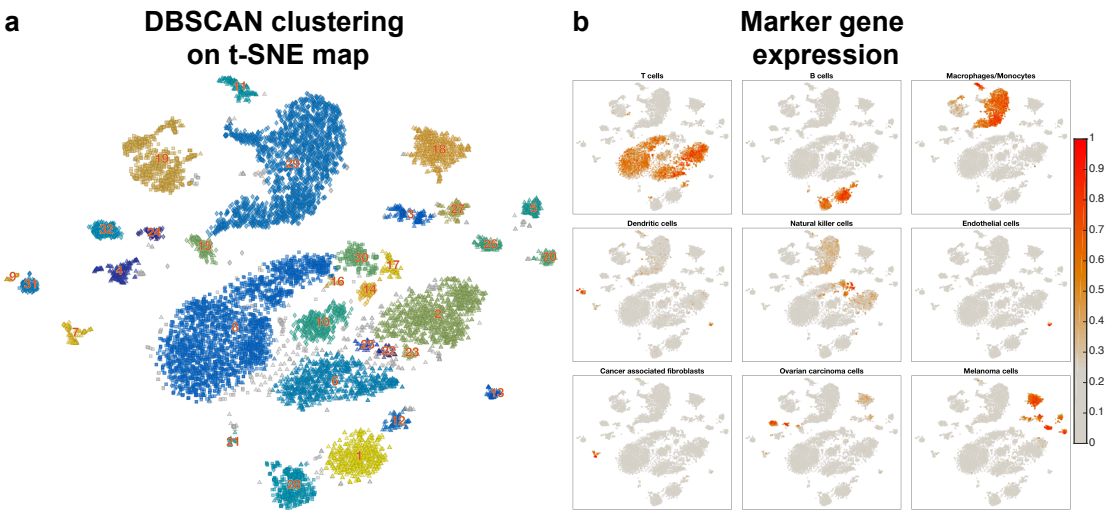

Fig. 5.13.: Clustering and marker gene expression of scRNA-seq data. (a) DBSCAN clustering was performed on the t-SNE map. Parameters of the algorithm were adjusted in such a way that distinct cell groups are identified as separate clusters. Symbols represent individual cells, shape of the symbol denotes indication. Colors and numbering denote the resulting clusters. **(b)** Heatmap of marker gene expression on the t-SNE mapping. For each cell type, the expression of cell type-specific marker genes was evaluated and merged to a score using three logical operators (AND, OR, NOT). Resulting scores show mostly distinct populations of cell types.

genes that are all required, (2) OR genes where only the expression of one of them is necessary, and (3) NOT genes where the expression is a negative selection criterion. The marker genes for each operator and cell type are indicated in Table 5.4.

The combined cell type score was normalized to $[0, 1]$ and plotted as a heatmap on top of the t-SNE map as shown in Figure 5.13b. For each cluster, the predominant cell types were determined by keeping only those cell populations that have a total score of at least $75\,\%$ of the maximal score. This helps to prevent misclassification of closely related cell types (e.g. Natural killer cells and T cells).

Cell type classification using decision trees

The resulting cell type assignment was used as a training set for a decision tree classification approach (see also Section 2.2.5). The t-SNE representation of the training set is depicted in Figure 5.14a.

The decision tree classifier, as shown in Figure 5.14b, performs several binary decisions based on the expression value of individual features in the training set. The accuracy of the classifier was evaluated using 5-fold cross-validation indicating a high classification accuracy ($98.06\,\%$). Subsequently, the trained classifier can be utilized to predict the cell

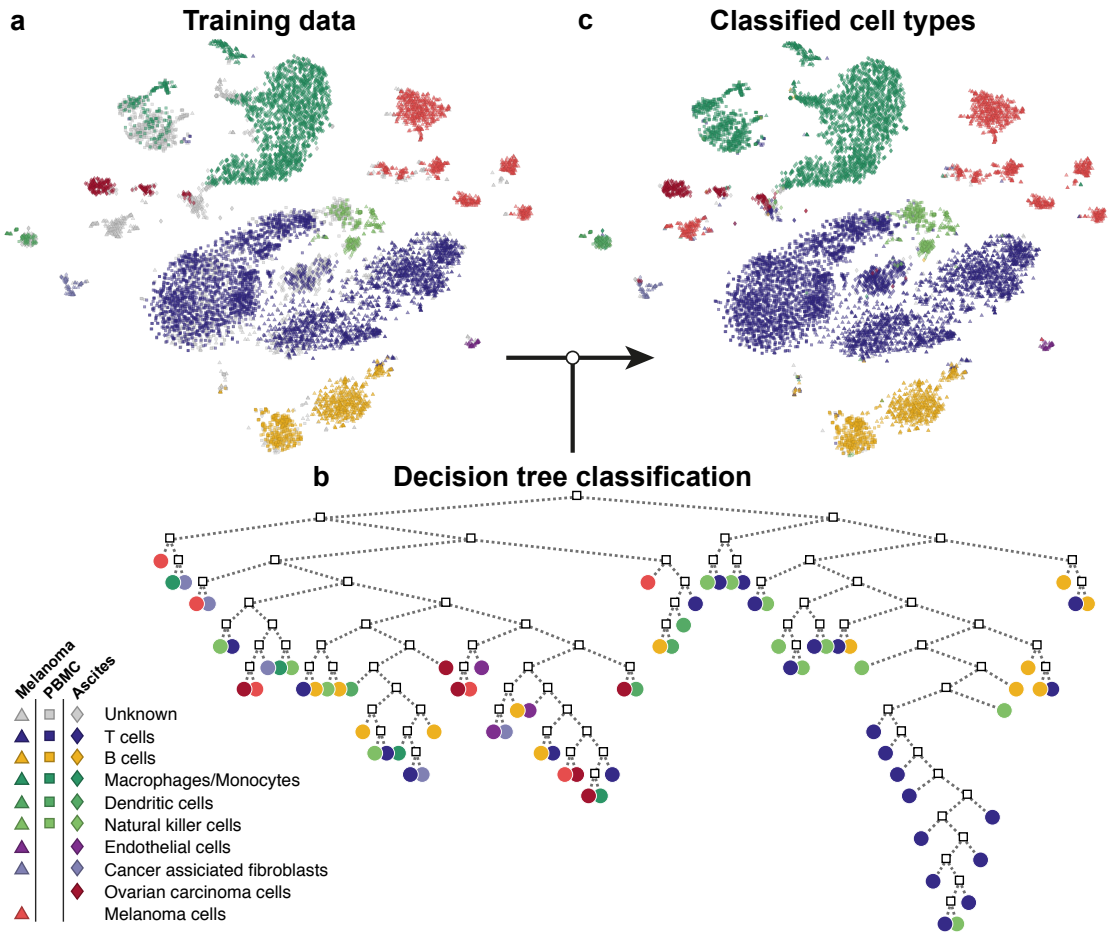

Fig. 5.14.: Identification of cell types of scRNA-seq data using decision tree classification. **(a)** t-SNE mapping of the training set. Predominant cell populations within each cluster were identified with the corresponding cell type and were used as training set for classification. 8864 of 11 759 cells were used for training. **(b)** Decision tree classification of nine major cell types. Each node in the graph represents a binary decision based on the value of one feature. Expression profiles were reduced to the *merged* gene set and transformed using PCA before training the classifier. **(c)** Predicted cell types upon decision tree classification. Cells with a posterior probability of less than 0.99 were labeled as "Unknown". After classification, 11 449 of 11 759 cells were assigned to one of the nine major cell types. **In all panels:** Symbols represent individual cells, shape of the symbol denotes indication, color indicates cell type according to the legend.

Tab. 5.4.: Marker genes used for cell type identification. For nine major cell types and three T cell subtypes, marker genes were defined based on suggestions from literature. Each gene was assigned to one of three logical operators (AND, OR, NOT) depending on its known expression in the respective cell type.

Cell type	AND marker	OR marker	NOT marker
T cells	CD3D, CD3E, CD3G, CD27, CD28		
CD4+ T cells	CD4		FOXP3, IL2RA, CTLA4
CD8+ T cells	CD8B	CD8A	CD4
Regulatory T cells	CD4, FOXP3, IL2RA, CTLA4		
B cell	CD19, MS4A1, CD79A, CD79B, BLNK		
Macrophages/ Monocyte	CD14, CD68, CD163, CSF1R, FCGR3A		
Dendritic cell	IL3RA, CLEC4C, NRP1		
Natural killer cells	FCGR3A, FCGR3B, NCAM1, KLRB1, KLRB1, KLRC1, KLRD1, KLRF1, KLRK1		
Endothelial cell	VWF, CDH5, SELE		
Cancer associated fibroblasts	FAP, THY1, COL1A1, COL3A1		
Ovarian carcinoma cells	WFDC2, EPCAM, MCAM		
Melanoma cell	PMEL, MLANA, TYR, MITF		

types of the whole data set, leading to the classification result shown in Figure 5.14c. Cells with a posterior probability lower than 0.99 were marked as “Unknown”.

Separate identification of T cell subtypes

Because of the strong similarity among T cell subtypes, the first round of classification was only performed on the nine major cell types indicated in Figure 5.13 and 5.14. Identification of three selected T cell subtypes, i.e. CD4+, CD8+ and regulatory T cells, was achieved by a second round of classification, taking only the previously determined population of T cells into account. The parameters for DBSCAN clustering needed to be adjusted to `MinPts=25` and `Eps=1.75` and the cross-validation showed an accuracy of $93.88\,\%$. The subtype classification is depicted in Figure 5.15.

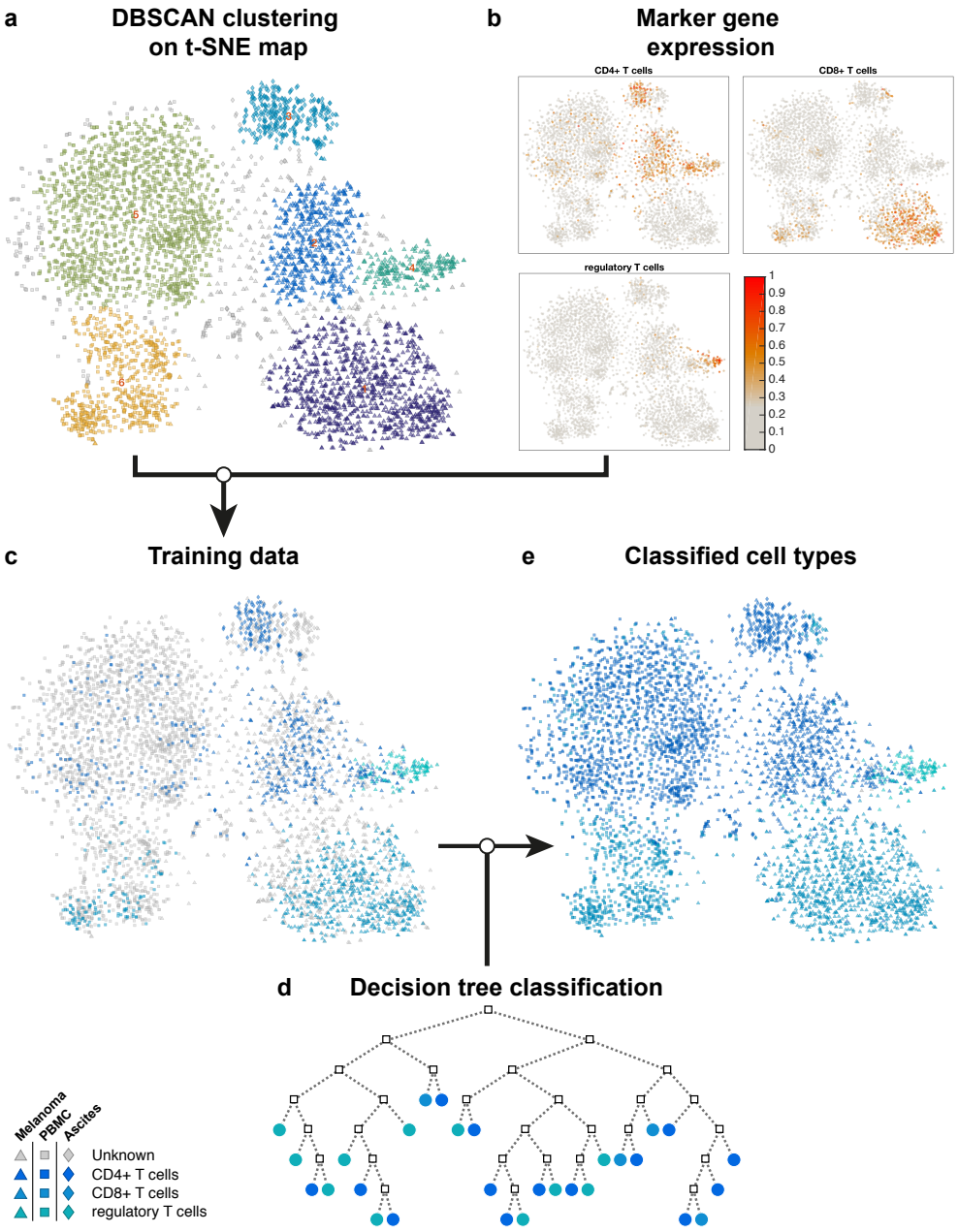

Fig. 5.15.: Classification of three T cell subtypes. **(a)** DBSCAN clustering on t-SNE map of the T cell population. Colors and numbering denote the resulting clusters. **(b)** Heatmap of marker gene expression on the t-SNE mapping. Expression of subtype-specific marker genes was evaluated and merged to a score using three logical operators (AND, OR, NOT). Resulting scores show mostly distinct populations of T cell subtypes. **(c)** t-SNE mapping of the training set. Predominant subtype populations within each cluster were identified with the corresponding T cell subtype and were used as training set for classification. 1439 of 5067 cells were used for training. **(d)** Decision tree classification of three T cell subtypes. Each node in the graph represents a binary decision based on the value of one feature. Expression profiles were reduced to the *merged* gene set and transformed using PCA before training the classifier. **(e)** Predicted subtypes upon decision tree classification. Cells with a posterior probability of less than 0.99 were labeled as "Unknown". After classification, 5052 of 5067 cells were assigned to one of the three T cell subtypes.

5.3.4 Analyzing indication and patient-specific gene expression

Effects of the tumor micro-environment

The final classification result is shown in Figure 5.16a. Cell types and T cell subtypes are indicated by different colors, source tissues are distinguished by the shape of the marker symbol. In the t-SNE map, we find distinct clusters of cells according to the cell type. Within each cell type cluster, sub populations corresponding to the source tissue could be identified, indicating that immune cells potentially change their gene expression depending on their micro-environment.

The similarity of corresponding cell types from different sources was further analyzed by calculating the pair-wise Pearson's correlation of the averaged gene expression profiles for each cell type. The resulting dendrogram of the correlations is shown in Figure 5.16b. In accordance with the t-SNE mapping, we find a high similarity ($\rho > 0.7$) for T cell subtypes originating from different sources. For natural killer cells, a correlation of $\rho > 0.8$ was observed for all three indications. While B cells from melanoma and blood samples are highly similar ($\rho > 0.9$), those from ascites samples are closer related to the respective dendritic cells ($\rho = 0.78$) than to the other B cells ($\rho = 0.7$). Malignant cells from ovarian carcinoma and melanoma are less correlated with each other ($\rho = 0.54$) than with tumor-associated cells ($\rho > 0.56$). Interestingly, despite their different origin, cancer associated fibroblasts from melanoma and ascites are highly similar in their gene expression ($\rho = 0.81$).

The cellular composition of each patient sample is shown in Figure 5.16c. For the 19 melanoma samples, the composition covers the full range from no tumor at all in sample 58 up to a very high tumor content for sample 78. For all four ascites samples, a huge fraction of macrophages/monocytes was identified while for PBMCs from blood samples T cells represent the biggest fraction.

The classification result was also compared to the original classification of the published melanoma data set by Tirosh *et al.* [190] and, for the ovarian ascites data set, to FACS data of three of the samples. As depicted in Figure 5.17 the classification is in line with previously published results and with the experimental measurements.

Malignant cells exhibit patient-specific gene expression

By mapping the sample/patient ID to the t-SNE map, as shown in Figure 5.18, we could confirm previous observations by Tirosh *et al.* [190] that malignant cells cluster mostly in a patient-specific manner. By contrast, for immune cells, heterogeneous clusters for

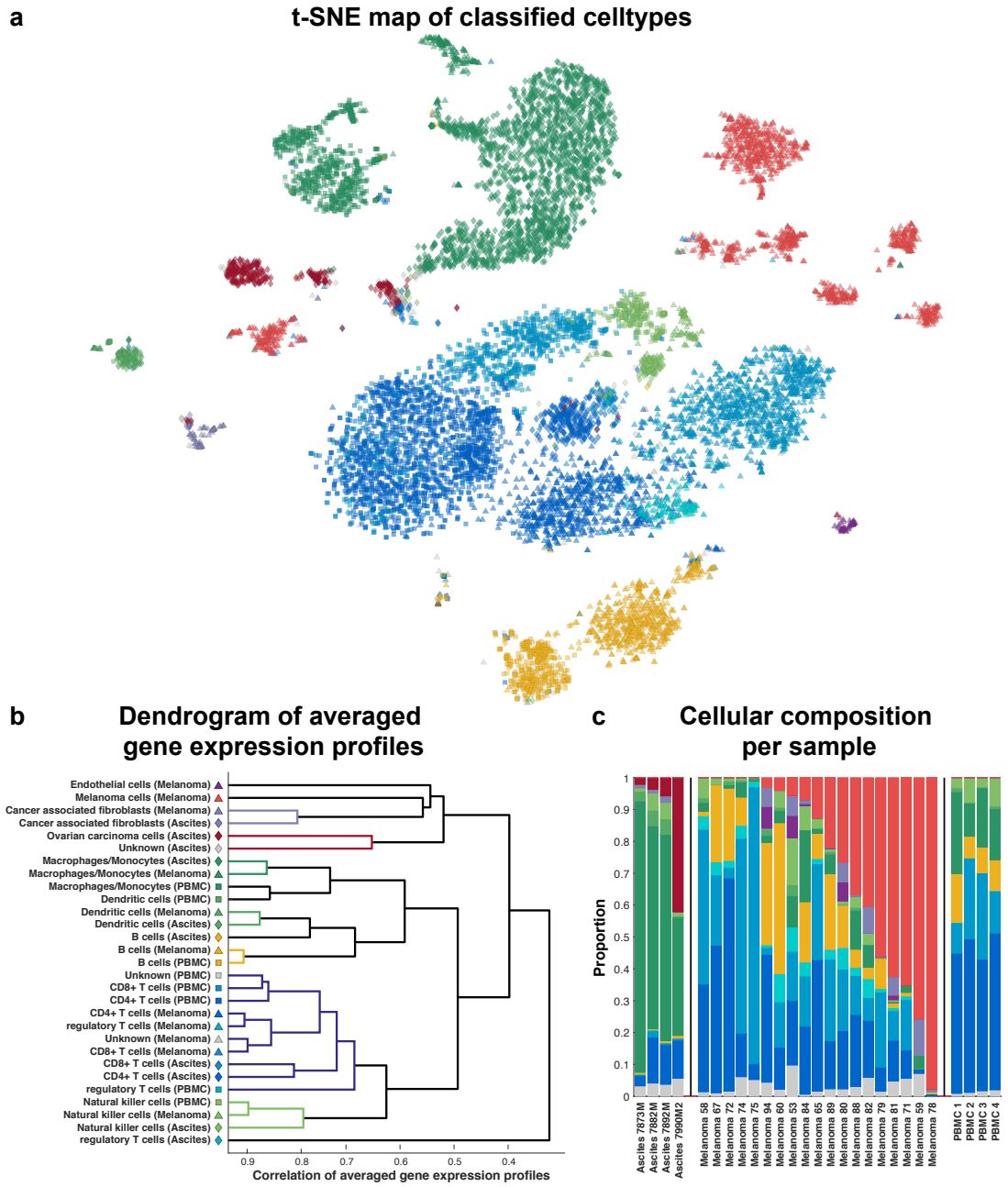

Fig. 5.16.: t-SNE map of single-cell data from three sources. (a) Single cells (symbols) were arranged in two dimensions based on similarity of their gene expression profiles by a dimensionality reduction technique (t-SNE). The clusters that emerge spontaneously can be associated with cell types (colors) and source location (symbol types: triangles for melanoma, squares for PBMCs, and diamonds for ascites; legend as in panel b). **(b)** Pair-wise correlation of averaged gene expression profiles of clusters encoding cell type and source location as identified in a) visualized as dendrogram. **(c)** Composition of individual tissue samples broken down by cell types.

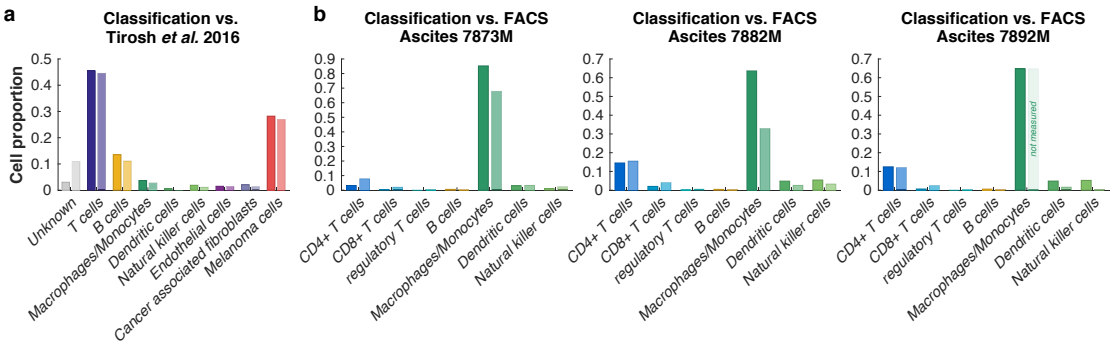

Fig. 5.17.: Comparison of the classification result with literature and experimental FACS analysis. (a) The result of our cell type classification (left bars, dark colors) compared to the cell types provided across all melanoma samples in the data set by Tirosh *et al.* [190] (right bars, light colors). **(b)** Cell type classification (left bars, dark colors) compared to FACS data (right bars, light colors) for three ovarian ascites patient samples. For sample 7892M, macrophages/monocytes could not be detected by FACS.

each cell type rather than for each patient sample were observed. Within each cell type cluster, sub-clusters are formed according to the tissue source.

5.3.5 Defining a benchmark based on single-cell gene expression data

After having identified patient- and indication-specific gene expression profiles, we want to benchmark the deconvolution approach depending on the information provided to the algorithm.

Generating bulk patient samples from single-cell data

By using single-cell gene expression data, we can construct "bulk" gene expression profiles for each of the 27 patient samples by aggregating all single-cell gene expression data. For these artificial samples, the actual composition is known *a priori* and serves as the ground truth for benchmarking the deconvolution approach.

Defining reference gene expression profiles

To study how the results of deconvolution of bulk expression data are affected by micro-environment-specific changes in gene expression and by patient-to-patient variation, we consider three configurations of *reference gene expression profiles* (RGEPs). The construction process is schematically depicted in Figure 5.19. Starting from several patients for each indication (for the sake of convenience, in the example we consider

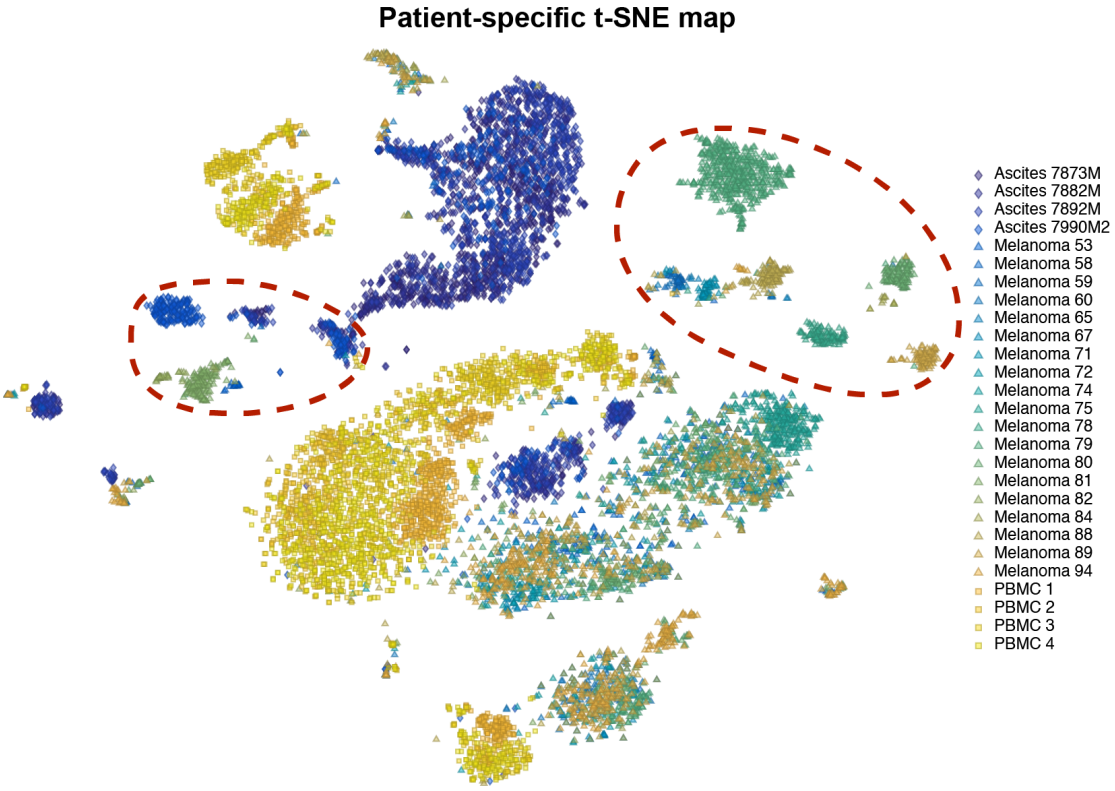

Fig. 5.18.: Patient-specific t-SNE map of scRNA-seq data. Malignant cells (highlighted by red dashed line) cluster according to patient sample while immune cells cluster primarily according to the cell type. Symbols represent individual cells, shape of the symbol denotes indication, color indicates patient of origin according to legend.

only two patients per indication), we define *RGEP1* as the population average over all single-cell gene expression profiles of a certain cell type originating from whole blood samples. Because there are no tumor associated cell types contained in the blood samples of healthy donors, the reference profiles for tumor cells, endothelial cells and cancer associated fibroblast are unknown. For *RGEP2* we use the population average of the expression profiles from all indications. As stated previously, tumor cells from melanoma and ascites can be very different in their gene expression and thus require separate reference profiles. The third configuration, *RGEP3*, averages only over expression profiles of cells belonging to the same indication. The indication-specificity might be necessary as corresponding cell types from different sources can show altered expression characteristics.

In addition to the three RGEPS, we set up two artificial control scenarios that are extensions of RGEP3 and include patient-specific information. These scenarios are not applicable in the real world, but serve to evaluate the relative importance of patient-specific information. In *control scenario 1* (CNTR1), patient-specific profiles are used only for the malignant cells, whereas consensus profiles are used for each non-malignant

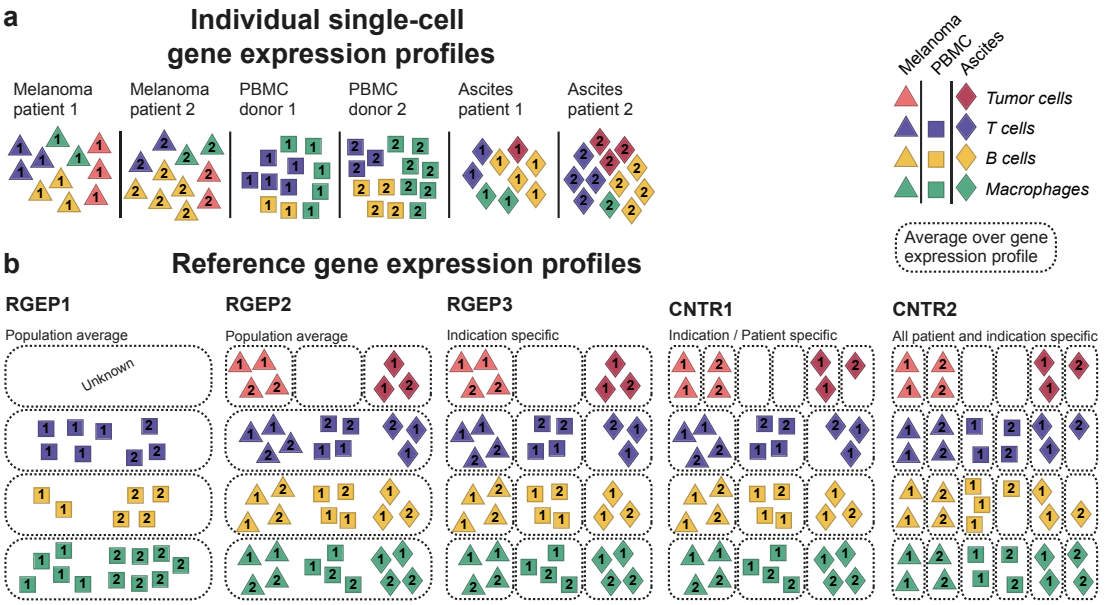

Fig. 5.19.: Generating cell type-specific gene expression profiles from single-cell data. (a) Individual single-cell expression profiles are obtained from scRNA-seq and illustrated by cell type (colors) and patient (numbers). **(b)** Five different reference profile configurations: (REGP1) Unknown tumor profile, profiles of normal cells obtained by averaging over all cells of a certain cell type from PBMCs; (REGP2) reference profiles are obtained from scRNA-seq data from three indications and averaged over all patient samples. Tumor cells are indication-specific; (REGP3) All profiles are indication-specific; (CNTR1) immune and stromal cell profiles are indication-specific while tumor profiles are patient-specific; (CNTR2) All profiles are patient-specific.

cell type. In *CNTR2*, patient-specific profiles are used for all cell types. In principle, CNTR2 serves as the upper limit on what is technically possible using deconvolution approaches.

Signature gene sets for deconvolution

Besides the reference gene expression profiles, we need to define which genes we include for the deconvolution. Based on published lists of characteristic genes from literature, we define five *signature gene sets* and evaluate their impact on the deconvolution accuracy. The smallest one is taken from *Table S12* of the supplementary information of Tirosh *et al.* [190]. It consists of 244 genes that were shown to be differentially expressed in populations of T cell subtypes. Thereof, 239 are contained in the PBMC, melanoma and ascites data sets. The second gene set is obtained from *Table S3* of Tirosh *et al.* [190]) and represents a list of 391 genes that have been identified as differentially expressed among the cell types. Thereof, 374 genes are contained in all three scRNA-seq data sets. The *LM22* gene set by Newman *et al.* [139] consists of 547 genes of which 496 are contained

in our data sets. The *Merged* gene set is generated by merging all genes from the *LM22*, *Table S3* and *Table S12* gene sets and adding the 45 marker genes used for classification training. It consists of 1076 unique genes, thereof 1015 genes are in common with the scRNA-seq data. As a last gene set, *All genes* $17\,936$ genes that are contained in the all three scRNA-seq data sets are used. Thereof, $17\,933$ have a non-zero expression for at least one single-cell profile.

Deconvolution algorithms used for benchmarking

Next, we included a set of four algorithms to solve the deconvolution problem defined in (5.32): As introduced in the CIBERSORT method, ν-Support vector regression (ν-SVR) can be used. Here, we utilized the implementation of libSVM [24] for MATLAB (version R2016a, The MathWorks Inc., Natick, MA, USA). The parameters were set to `-s 4 -t 0 -n 0.50 -h 0 -c 1 -q`. The `mldivide` function from MATLAB uses the pseudo-inverse of the matrix B for solving for `w = pinv(B)*m`. This is equivalent to using `w = mldivide(B, m)`. The `fitlm` function from MATLAB fits a linear model to the data based on a least-squares fit. The main difference to the `mldivide` function is that for `fitlm` an intercept is taken into account. For the CVX package [72] for MATLAB the problem was defined as:

```
cvx_begin quiet
    cvx_solver sdpt3;
    variable w(size(B, 2)) nonnegative;
    minimize( norm((B*w - m), 2) + lambda*norm(f, 2))
    subject to
       w <= 1;
       sum(w) <= 1;
cvx_end
```

with `lambda=1` and solved using the SDPT3 algorithm [191] for semidefinite-quadratic-linear programming problems.

Processing of estimation results

The results for the proportions of known cell types w, as obtained by one of the above-mentioned algorithms, are processed by replacing negative numbers by zeros [139]. The proportion of unknown other cell types $\tilde{w}$, i.e. cell types for which no reference profile

was available, is calculated by taking the difference between one and the sum of all n known cell proportions:

$$\tilde{w} = 1 - \sum_{i=1}^{n} w_i \tag{5.35}$$

Assessing the estimation accuracy

To assess the accuracy of our deconvolution results, we compared the true cell fractions, as calculated from the number of single-cell measurements for each cell type and patient, with the estimation result by calculating Pearson's correlation coefficient ρ for all patients. We quantified the uncertainty of our quality measure by performing bootstrap re-sampling (100 replications) of our deconvolution results and calculated the median and lower and upper quartiles.

5.3.6 Impact of signature gene sets on deconvolution accuracy

We start evaluating the performance of different gene sets when deconvolving the generate bulk samples using the ν-SVR algorithm. The reference gene expression profiles were constructed using the RGEP3 configuration. As shown in Figure 5.20a, the overall accuracy is relatively high for all gene sets ($\rho > 0.94$). However, the *Merged* and *All genes* gene sets perform best with an overall correlation of $\rho > 0.97$. Looking at the performance for individual cell types, we find only minor differences between the five gene sets. Interestingly, despite the slightly lower overall correlation, the *Merged* gene set outperforms *All genes* for several cell types such as natural killer cells, dendritic cells and T and B cells. Furthermore, due to the much smaller size of this gene set (1015 vs. $17\,933$), the computation time using the *Merged* gene set is by a factor of 123 smaller ($13\,\mathrm{s}$ vs. $1600\,\mathrm{s}$ on an Intel Xeon X5680 $3.33\,\mathrm{GHz}$ CPU with $94\,\mathrm{GiB}$ RAM). We therefore pick the *Merged* gene set as the best performer for all further analyses.

5.3.7 Evaluating deconvolution algorithms

In the next step, we analyze the impact of the chosen algorithm on the deconvolution performance. We use the *Merged* gene set and the RGEP3 configuration. The result is depicted in Figure 5.20b. Again, the overall performance is very high ($\rho > 0.97$) for all four algorithms but the highest for `filtm` and ν-SVR ($\rho > 0.98$). While `filtm` performs slightly better for melanoma cells ($\rho = 0.99$ vs. $\rho = 0.97$), ν-SVR performs better for all other cell types. Thus, we can confirm the accuracy achieved with the ν-SVR algorithm

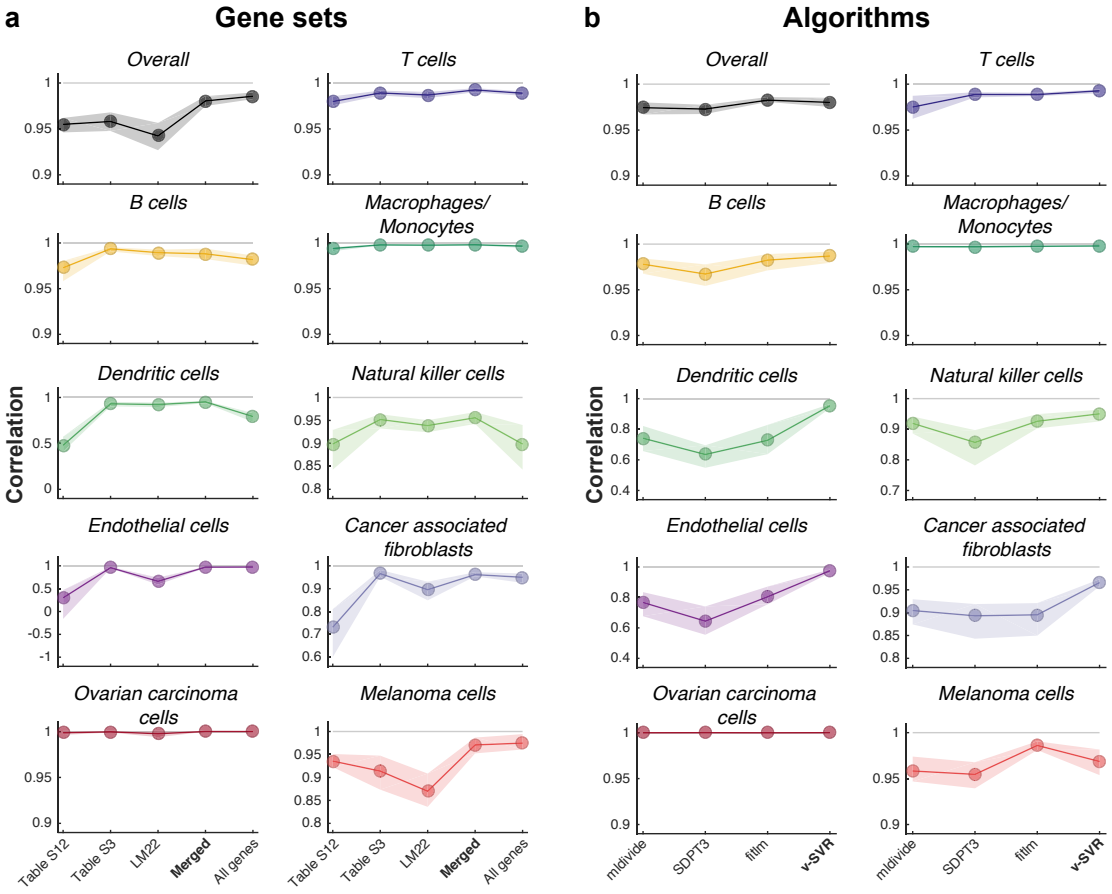

Fig. 5.20.: Comparison of different gene sets and algorithms. **(a)** The deconvolution accuracy of five different gene sets was assessed for the nine major cell types based on the ν-SVR algorithm. **(b)** Four deconvolution algorithms are compared based on the deconvolution result obtained with the *merged* gene set. Each dot represents the correlation of the ground truth with the deconvolution result over all patient samples. The shading shows the upper and lower quartile obtained by bootstrapping. Only the results for RGEP3 are indicated. The detailed scatter plots for all configurations are shown in Figure B.9–B.15 in the appendix.

as originally observed with the CIBERSORT method on microarray data and select this algorithm for all subsequent analyses.

5.3.8 Impact of patient- and indication-specific reference profiles

Having selected the best performing signature gene set and deconvolution algorithm we now want to go into greater detail of the deconvolution result and analyze the role of patient- and indication-specific reference profiles. Therefore, for each of the five reference profile configurations (RGEP1-3, CNTR1-2), we run the deconvolution and benchmark the result against the ground truth. When plotting the estimated proportion against the

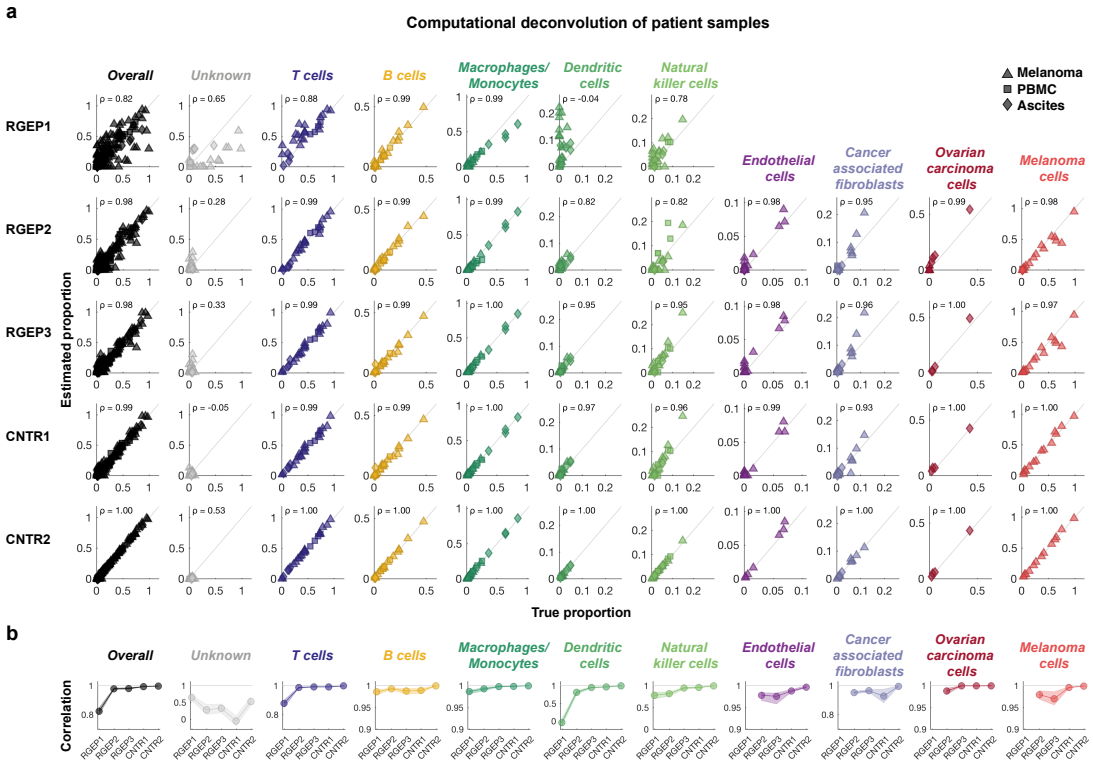

Fig. 5.21.: Benchmark result for five RGEPs and nine major cell types. (a) Scatter plot of true and estimated cell proportions for all 27 patient samples. Each dot represents one patient sample. Values close to the diagonal correspond to high deconvolution accuracy. Columns depict cell types; rows describe the five different configurations (REGP1-3 and CNTR1-2). ρ denotes the Pearson's correlation coefficient. In configuration REGP1, estimates for tumor-associated cell types are not available. (b) Pearson's correlation coefficient between estimated and true cell fraction for all five configurations. Dots denote the median of the correlation coefficient; the shading represents the uncertainty based on bootstrapping (upper and lower quartile). (Please note the different scaling of the figure axes.)

true proportion, as shown in Figure 5.21a, we can determine the correlation over all patient samples. Each row represents one configuration of reference gene expression profiles while the columns represent cell types.

For RGEP1, the fractions of unknown cells are relatively high as no reference profiles were available for tumor associated cell types. Nevertheless, a mediocre overall result ($\rho = 0.82$) can be achieved ranging from very weak results for dendritic cells ($\rho = -0.04$) up to more or less exact deconvolution fractions for B cells and macrophages/monocytes ($\rho = 0.99$). In RGEP2, melanoma and ascites data are taken into account providing reference profiles for endothelial cells, cancer associated fibroblasts and ovarian carcinoma as well as melanoma cells. The overall performance increased significantly to $\rho = 0.98$ reflecting the very high accuracy for most cell types ($\rho > 0.94$). Only for dendritic cells

and natural killer cells, despite the increase as compared to RGEP1, the accuracy is still moderate ($\rho = 0.82$). As the proportion of unknown cells is very small (see also Figure 5.16c), the correlation is no longer a good measure of accuracy. Adding indication-specific reference profiles for all cell types in RGEP3 further improves the deconvolution accuracy. Estimates for all cell types reach correlations of $\rho > 0.95$ indicating that natural killer cells and dendritic cells require indication-specific information for an accurate and reliable deconvolution. When adding patient-specific information for the tumor cells in CNTR1 and for all cell types in CNTR2, the accuracy can be improved even more until correlations of $\rho = 1.00$ are reached for CNTR2. These reference profiles, however, are not available in real world scenarios and serve only as an upper limit of the methodology.

In summary, our analysis revealed that reference gene expression profiles obtained only from PBMC data are insufficient to reliably deconvolve patient samples from other indications. Adding data from multiple indications could improve the accuracy for all cell types. However, certain cell types such as dendritic cells and natural killer cells seem to require indication-specific profiles for an accurate estimation of the sample composition.

5.3.9 Estimating T cell subtypes and prognostic ratios

So far, we focused our analysis to nine major immune and tumor-associated cell types. Our classification of scRNA-seq data also allows to distinguish between multiple T cell subtypes, namely CD4+, CD8+ and regulatory T cells. Quantification of T cell subtypes is especially interesting since it was shown that location and abundance can be used to predict patient outcome [64, 90]. As in the previous section, we perform the deconvolution on the *Merged* gene set using the ν-SVR algorithm and evaluate all five configurations of reference gene expression profiles. The resulting scatter plots are shown in Figure 5.22a.

While for CD8+ T cells, the estimation results are accurate ($\rho > 0.94$) for all five settings, for CD4+ and regulatory T cells, the estimation results in RGEP1 are only mediocre ($\rho = 0.63$ and $\rho = 0.43$) but improve significantly for RGEP3 ($\rho = 0.87$ and $\rho = 0.94$). This is also reflected in the ratios of Treg/CD4+, CD8+/Treg and CD4+/CD8+ T cells that reach accurate estimations for RGEP2 ($\rho = 0.94$, $\rho = 0.96$ and $\rho = 0.93$). The estimation for all T cell subsets and ratios does not significantly improve for CNTR1 but does improve in CNTR2 ($\rho = 1.00$), indicating that gene expression of T cells is influenced by patient-specific effects.

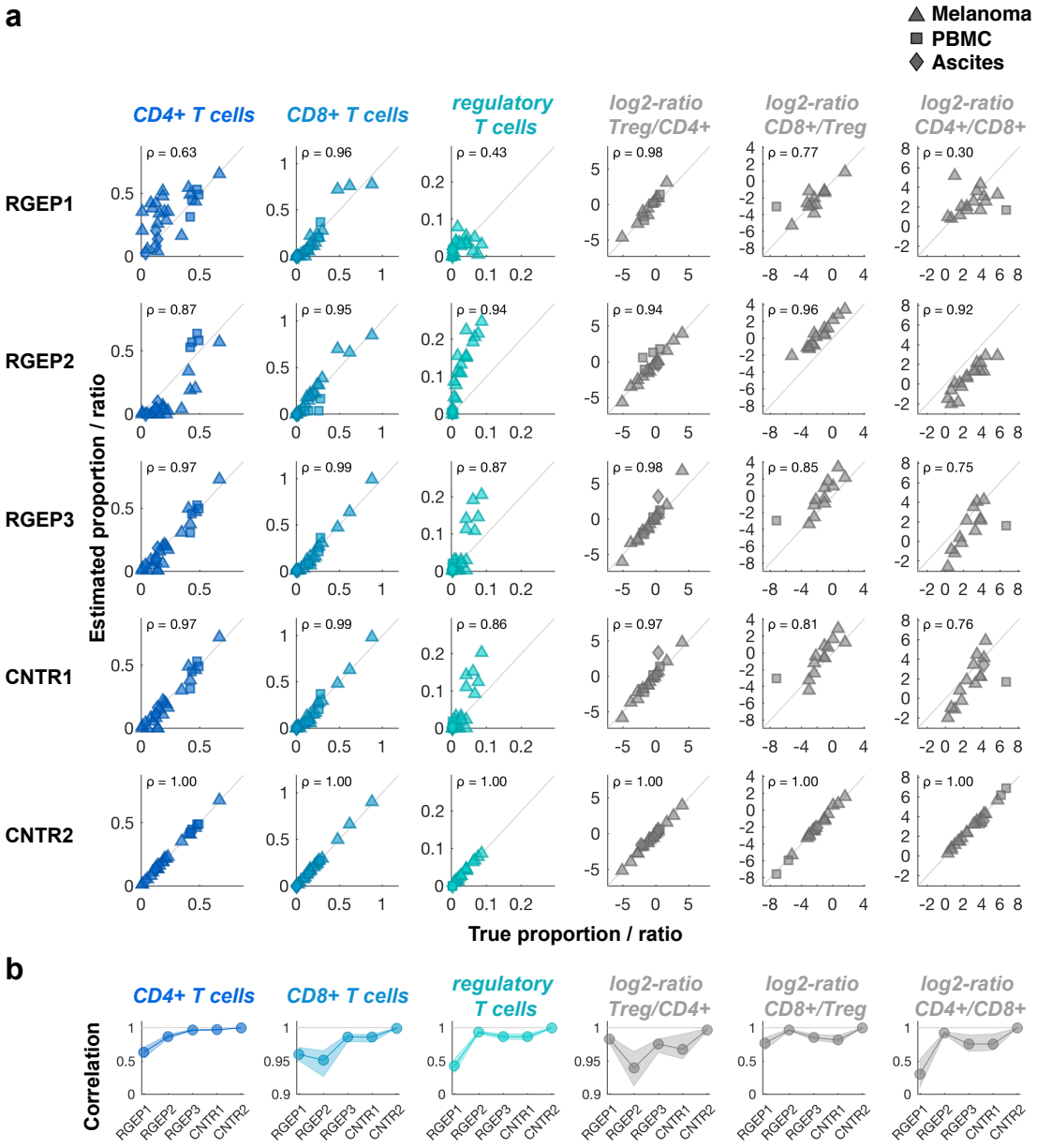

Fig. 5.22.: Benchmark result for five RGEPs and three T cell subtypes and three T cell ratios. Caption as for Figure 5.21, ratios are indicated on a $\log_2$-scale.

In summary, deconvolution using gene expression profiles based on data from the respective micro-environment (RGEP3) were sufficient to obtain reliable estimates of the cellular composition of the samples. Deconvolution using gene expression profiles based on data from peripheral blood (RGEP1) were considerably less accurate.

5.3.10 Predicting tumor expression profiles

As shown in Figure 5.18, the gene expression tumor cells clusters in a patient-specific manner. In a real word scenario, however, it is unlikely that the tumor gene expression can be quantified separately for each patient as only the bulk sample expression is measured. Based on our deconvolution result obtained with the indication-specific RGEP3, we can predict the tumor cell gene expression for each patient individually. Therefore, we simply need to subtract the weighted expression of non-tumor cells from the bulk sample expression and rescale the expression with the estimated percentage of tumor cells, i.e.

$$\vec{t}_i = \frac{\vec{m}_i - B_{\mathrm{non-tumor}} \cdot \vec{w}_{i,\mathrm{non-tumor}}}{w_{i,\mathrm{tumor}}}, \tag{5.36}$$

where $\vec{m}_i$ represents the bulk sample expression of the i-th patient, $B_{\mathrm{non-tumor}}$ is the signature matrix for all non-tumor cell types and $\vec{w}_{i,\mathrm{non-tumor}}$ and $w_{i,\mathrm{tumor}}$ are the non-tumor and tumor cell fractions of the patient sample.

In order to evaluate the accuracy of the predicted tumor profiles, we analyze the correlation of the predicted profile with the true expression profile of the tumor cells. Additionally, we want to know whether the estimated profile is reflecting the true tumor gene expression better than the expression of the whole bulk sample. As a baseline, we also plot the expression of the non-tumor cells in each sample against the tumor. The scatter plots of the comparison are depicted in Figure 5.23a. The patient samples are sorted by the estimated percentage of tumor cells and Pearson's correlation ρ is indicated in the respective color. Clearly, for samples with low tumor cell content, our prediction is less accurate ($0.39 \leq \rho \leq 0.69$) than correlating the bulk sample with the tumor ($0.75 \leq \rho \leq 0.82$). When exceeding an estimated percentage of tumor cells of $20\,\%$, however, the estimated tumor profile is improved and outperforms the correlation with the bulk sample ($0.93 \leq \rho \leq 0.99$ vs. $0.86 \leq \rho \leq 0.98$). For the upper end, i.e. when the sample consists of more than $70\,\%$ of tumor cells, the correlation of bulk sample and our prediction with the true tumor cell profile converge to correlation values of $\rho \geq 0.99$.

The dependence of the prediction accuracy on the percentage of tumor cells in the sample is shown in Figure 5.23b. For low tumor content, non-tumor and bulk sample gene expression are very similar and result in the same basal correlation of 0.7 to 0.8. In this range, our prediction for the tumor profile performs worse because the low signal-to-noise ratio. For more than $20\,\%$ tumor content in the sample, we can improve the prediction of the gene expression significantly until the bulk sample correlation converges for samples with more than $70\,\%$ tumor content. The baseline correlation of non-tumor

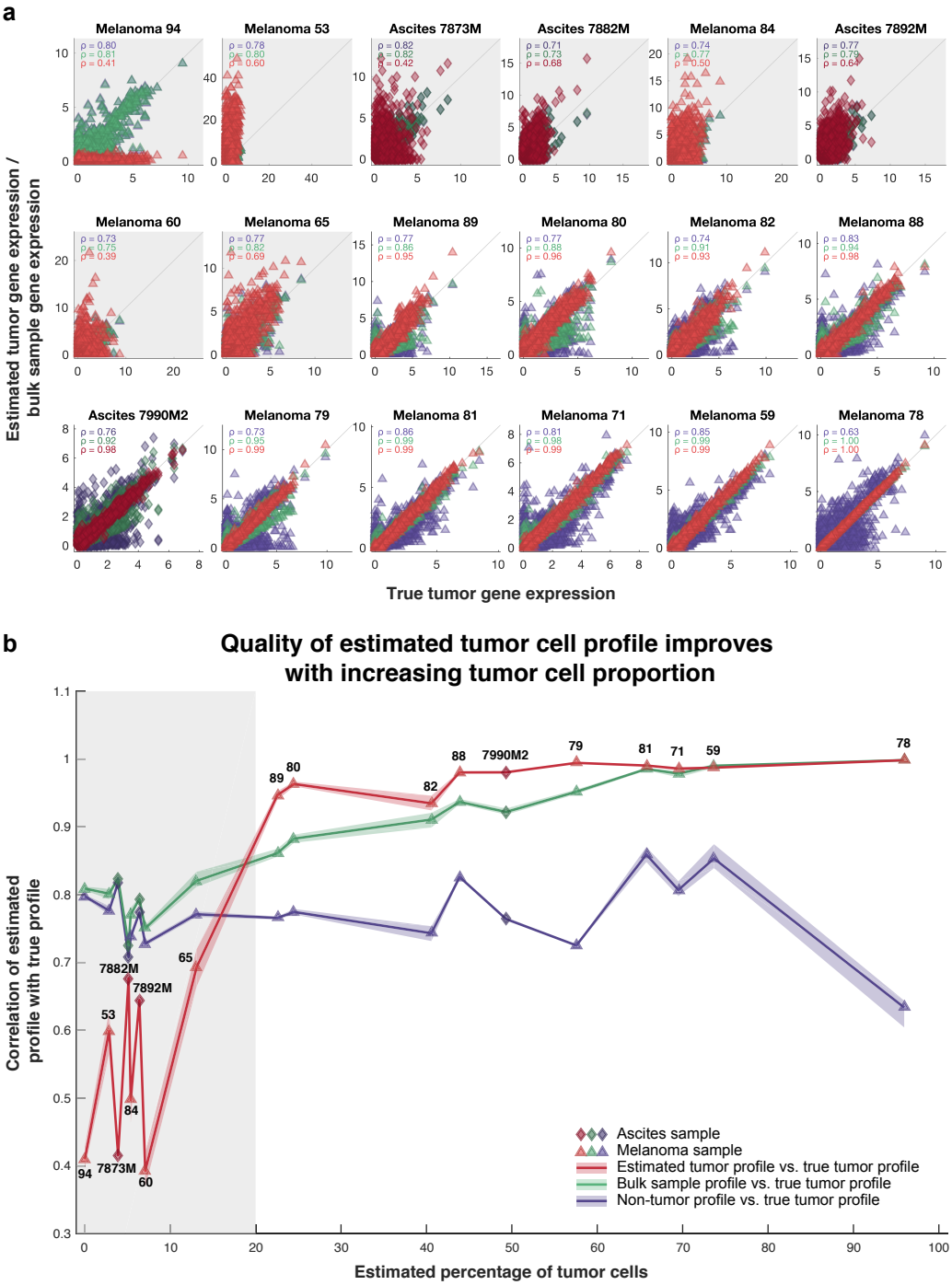

Fig. 5.23.: Estimation accuracy of patient-specific tumor cell gene expression profiles. (a) Scatter plot of predicted vs. true transcriptome wide gene expression (17 933 genes) of the tumor cells. Patient samples without any tumor cells have been excluded from this analysis. ρ denotes the Pearson's correlation coefficient including uncertainty based on bootstrapping. **(b)** Correlation values from (a) plotted against the estimated proportion of tumor cells for each patient sample. Symbols and numbering denote individual patient samples.

vs. tumor profile, by contrast, drops for samples with high tumor content as less and less non-tumor cells are left in the sample.

In summary, for patient samples with a tumor cell content of $20\,\%$ to $70\,\%$ the tumor cell gene expression profiles obtained using deconvolution are significantly improved as compared to the bulk sample profiles.

5.3.11 Discussion

The recent advent of single-cell sequencing technologies enables, for the first time, the in depth analysis of cellular heterogeneity of patient samples. However, to better understand how the interaction of different immune cell types contributes to fight malignant cells, it is essential to decipher the cellular composition of complex tissue samples. In the clinical routine, bulk gene expression of samples from tumor biopsies is measured using microarray or, more recently, using RNA-seq. While previous approaches to deconvolve the cellular composition from expression data mainly made use of reference profiles obtained by microarray of leukocytes and other PBMCs, the advantages of RNA-seq data were not yet exploited for this purpose.

With the presented approach we are able to accurately determine the composition of samples originating from 29 patients and three distinct tissue sources. As a proof-of-principle, we used bulk gene expression data constructed from single-cell measurements to benchmark the impact of several factors as the reference gene set, the deconvolution algorithm and the source and quality of the reference gene expression profiles. We showed that the choice of the genes used for deconvolution has only a minor impact on the accuracy of the deconvolution result. When using the indication-specific reference profiles (RGEP3), even with the T cell specific gene set *Table S12*, consisting only of 239 genes, the overall correlation of estimated vs. true cell fractions is still very high ($\rho = 0.95$) and the deconvolution only fails for dendritic cells, endothelial cells and cancer associated fibroblasts ($0.34 \leq \rho \leq 0.71$). A combination of multiple gene lists to the *Merged* gene set, consisting of 1015 genes, performed best and led to accurate predictions for all nine cell types ($\rho \leq 0.95$). The population of T cells could be further divided into three subtypes and prognostic ratios of the subtypes could be successfully determined with mediocre to high accuracy ($0.75 \leq \rho \leq 0.99$).

We also investigated the effect of the algorithm used for deconvolution. Again, the differences between the results obtained using different algorithms were relatively small and all four variants had a high overall accuracy ($\rho \leq 0.97$). The ν-SVR algorithm, as used by the CIBERSORT method, performed best for all cell types but the melanoma cells where the accuracy was still fairly high ($\rho = 0.97$).

The way how reference gene expression profiles are constructed, by contrast, strongly affects the deconvolution result. We investigated five variants with different data bases and increasing level of specificity in conjunction with the best performing gene set (*Merged*) and deconvolution algorithm (ν-SVR). While RGEP1, which is solely taking PBMC data into account, failed to estimate the fractions of dendritic cells ($\rho = -0.04$) and performed only mediocre for natural killer cells ($\rho = 0.78$), RGEP2-3 could significantly improve the accuracy. Thus, reference profiles obtained only from whole blood are insufficient to deconvolve samples from other tissue contexts, possibly altered by the tumor micro-environment. In RGEP3 source-specific reference profiles were generated resulting in a highly accurate deconvolution of sample compositions. As an upper limit, we also tested reference profiles using patient-specific information (CNTR1-2). These could further improve the results but are unrealistic for real world application.

Although the patients under investigation were suffering from the same malignancy (i.e. melanoma or ovarian carcinoma), the gene expression of malignant tumor cells was found to cluster by patient suggesting great differences between patients. Based on our estimation results we can "purify" the bulk expression profile and predict the gene expression of tumor cells on a patient-specific level. By correlating the predicted expression over all $17\,933$ genes contained in our data set with the true tumor cell expression obtained from the single-cell data, we could show that for a range of $20\,\%$ to $70\,\%$ tumor cell content in the sample, our prediction performs significantly better than correlating the bulk expression with the tumor.

A limitation of the presented approach can be found in the sequencing technology: While our analysis is based on single-cell RNA-seq data obtained by the newest generation of devices on the market, in clinical practice, biopsy samples from cancer patients are mostly analyzed by bulk sequencing often on older devices. Therefore, the compatibility of data obtained using bulk sequencing and single-cell technology needs to be systematically evaluated and possible conversion factors for gene expression would need to be defined. However, if sufficient data from both technologies are available for one and the same sample, this issue should be straight-forward to be resolved.

5.4 Conclusions

In this chapter, I demonstrated how computational methods from bioinformatics and systems biology can bring the diagnosis and treatment of cancer and associated health-conditions to the next level. On the way to a more personalized treatment of patients suffering from anemia, computational modeling becomes more and more a focus of publicly funded research, pharmaceutical industry and regulatory authorities. Severe side effects caused by chemotherapy can be reduced by adapting dosing regimens to patient-specific drug response and by supporting standard therapy with additional immunotherapeutical agents.

In parallel, increasing availability and decreasing costs for high-throughput sequencing technologies enable an integration into the clinical data acquisition pipeline. Together with large databases, such as TCGA and CCLE, these data will allow for extensive retrospective analyses and can lead to novel prognostic markers for specialized therapeutic approaches.

By adapting a published model of Epo-EpoR interaction and calibrating its parameters to *in vitro* data of ESA depletion in the supernatant of several human cell lines, a predictive tool for characterizing the properties of ligands and cells could be established. The extension of the model by a pharmaco-kinetic and -dynamic part allows for model-based risk prediction and for the development of an optimized dosing schedule on a patient-specific basis.

In the second part, an existing computational method for deconvolution of bulk tumor samples was adapted for the use with modern RNA sequencing data. Based on a benchmark study performed with single-cell RNA-seq data, we could show that only reference gene expression profiles from the tumor micro-environment allow for an accurate deconvolution. Therefore, indication specific reference profiles for each type of cancer need to be established in order to reliably predict the cellular composition of tumor samples.

Both projects presented in this chapter address clinical problems with methods from computational systems biology. Transferring results from basic research to the clinical context is one of the biggest challenges on the way to precision medicine. While our results could be validated *in vitro* and promising predictions for applications in patients could be made, a validation in clinics has yet to be performed in order to ensure safety and efficacy of the model predictions.

6 Conclusions & Outlook

In this last chapter, I want to summarize the main findings of the previous chapters and return to the questions raised in the introduction of this thesis. In addition, I will provide an outlook on the presented research topics and highlight the role of computational systems biology for the future development of treatment strategies.

6.1 Conclusions

In the first part of this thesis, I presented examples for community-driven software development for systems biology. By providing established methods for parameter estimation and identifiability analysis, these software tools enable researchers to construct and calibrate computational models of the systems of interest. Further, these software tools will help to use those models for generating valuable hypothesis and predictions.

In the subsequent chapters, I demonstrated how predictive computational models can utilized to foster understanding of intracellular processes during virus infection, to characterize ligand-receptor interactions and to optimize treatment in a patient-specific manner. Furthermore, based on computational deconvolution of high-throughput gene expression data, the infiltration of tumors by immune cells could be resolved which potentially enables the selection of adequate immunotherapy agents to support anti-cancer treatment.

Software development for systems biology

In systems biology, researchers from different disciplines such as molecular biology, physics and mathematics collaborate on joint efforts often spanning multiple universities or institutes. In order to make large data sets available to all partners within the project or to provide software tools to the scientific community, decentralized, web-based platforms show an increasing popularity.

In the scope of this thesis, I presented two projects: one for collaborative software development and one dealing with data availability for joint projects. For the first project we raised the following questions:

Can community-driven software development accelerate and improve workflows in computational biology research? What features do software solutions need to facilitate model construction, calibration and analysis for systems biology modeling?

In 2003, the standardized modeling language SBML was introduced. Since then, many software tools for model construction, simulation and parameter inference have been developed aiming to provide implementations of established methods and algorithms to the computational modeling community. The potential target audiences range from experimental biologists with very little experience in mathematical modeling to experts in the field of computational biology. As most of the applications have a non-commercial, academic background, the development is mostly driven by one or two research groups and the choice of the features is oriented towards the needs of their own research projects.

The software tool presented in this thesis, namely the *Data2Dynamics* (D2D) software, originated from a similar context but evolved differently: While the main focus still lies on efficient parameter estimation based on data obtained under multiple experimental conditions, the development is driven by a community of active researchers through the collaborative platform GitHub. Instead of one person in charge for the development, every user can suggest new features or implement a new version himself, which is then curated by the community of other users and developers. Therefore, the reliability as well as the extent of novel functionalities increased drastically within the last years. Thus we can conclude that the community-driven development approach helps to accelerate and improve workflows in computational biology research. The challenge of this concept lies in the structure and organization of the source code and demands mutual code revisions in order to keep the software running stably in the longer term.

In terms of the essential features of a software solution for systems biology modeling there are different aspects to be considered. First of all, SBML import and export capabilities are crucial to ensure compatibility of new software with existing tools and models. Other features depend very much on the target user base. Other than for instance COPASI, D2D has its main focus on data-based dynamic modeling which requires efficient implementations of parameter estimation and identifiability routines. On the other hand, rather than offering graphical user interfaces, D2D uses MATLAB scripts for the modeling workflow. While this restricts the target group to researchers with at least basic programming skills, it saves computational power for the actual tasks and renders complex simulation setups more transparent and reproducible. Taken together, SBML support and a lightweight and flexible architecture providing efficient algorithms for ODE integration and parameter estimation are the key features of D2D that facilitate model construction, calibration and analysis.

In the second project, a different set of questions was asked:

Which functionality enables the initial assessment and analysis of big data sets? How can data are made available to users from multiple institutions independently from their local infrastructure and technical background?

Modern high-throughput experiments generate huge amounts of data providing expression levels of individual genes or proteins for the whole cell under investigation. Data processing and analysis, as described in Section 2.2, can be challenging. High-dimensional data need to be made tangible for scientist in order to identify features of interest and to generate hypotheses. Within a collaborative research project, this is especially challenging as each group or institution mainly manages their own data sets.

In the presented case, a simple web-based platform was developed to provide all project partners with data acquired by contributing groups of the collaborative project. The aims were to make it easy to combine, browse and filter all available data giving researchers a first glimpse on the information contained therein. The backbone of the platform was written in the open-source programming language *R* which provides a variety of statistical analyses and tools for handling large data sets. The front-end, based on the package *R shiny*, provides a well structured and easy access to the data through the web-browser and is therefore independent of the operation system of the computer in use and does not require any programming knowledge. Hence, basic features such as merging, filtering and browsing multiple large data sets are very useful for a first assessment. Through a web-based platform, all project partners could access the data and run quick analyses directly on their local machines.

Identification of limiting factors during influenza virus infection

In this project, the influenza A virus infection was investigated. We focused on intracellular processes during the early infection and identified limiting factors during transport of the viral genome towards the host cell's nucleus. Our main questions were:

How can intracellular processes regulate host specificity and strain pathogenicity? Which nodes in the reaction network are limiting for the infection? How do changes in pH-sensitivity of the viral hemagglutinin affect the infectivity of IAV?

The infection of host cells with the influenza A virus requires a whole series of host-pathogen interactions in order to be successful. Starting with the entry of the virus into the host cell and the transport of the viral genome towards the nucleus, several factors could be identified that potentially modulate the host specificity and pathogenicity of a IAV strain. A critical step during this phase of the infection is the fusion of the

viral envelope with the surrounding endosomal membrane. The conformational change of the viral hemagglutinin protein is triggered by the acidification of the endosomal lumen. The pH value, at which virus-endosome fusion and therefore release of the viral genome takes place, is determined by the amino-acid sequence of HA. We showed that modifications of HA can lead to an altered pH sensitivity. In nature, variations of the HA segment occur due to antigenic shift or antigenic drift and were found to regulate the pathogenicity of different H5N1 mutants in ducks and chicken [165, 48]. Also, these mutations could be critical for adaptation of avian strains to mammals [216] and thus contribute to host-specificity. Moreover, HA was found to have lower pH threshold values for viruses from human isolates than for avian isolates of the same subtype [63]. By using a combination of deterministic and stochastic modeling approaches, we could show that a change in the pH-sensitivity of the HA protein results in an altered distance from the nucleus which has to be bridged by passive diffusion. In addition, we predicted that degradation of the viral genome during diffusion could limit infectivity which could be validated experimentally.

Therefore, we could improve the understanding on how host specificity and pathogenicity can be regulated by the characteristics of the HA protein. In the scope of this project, we took only the first steps of infection into account, neglecting subsequent processes like genome replication, synthesis of viral proteins as well as packaging, budding and release of progeny virions. Each of these steps relies on multiple host-factors as summarized e.g. in the FluMap [128] or in various extensive studies on the topic [101, 181, 208, 95]. Further computational studies on subsequent stages of infection require highly precise measurement techniques such as single-molecule FISH and super-resolution microscopy that allow observation of individual molecules within the infected cell over time.

Optimizing the treatment of cancer and associated health conditions

In the next chapter, I presented two projects on cancer-associated health conditions and showed how approaches from computational biology can not only enhance the understanding of involved intracellular processes, such as receptor-induced signaling, but also improve the *standard of care* treatment strategies.

In the first part of the chapter, we asked the questions:

Is there a better way to treat anemic patients? How can an optimal dosing schedule for individual patients be achieved?

Anemia can be treated either by blood transfusions or by stimulating the production of red blood cells using *erythropoiesis stimulating agents* (ESAs). The binding characteristics

of different ESAs determine the half-life in the body and therefore the dosing schedule in clinics. The response to the ESA treatment depends on the amount of susceptible progenitor cells within the patient's bone marrow.

Here, an existing model of ligand-receptor interaction was used to characterize novel therapeutic ligands based on the depletion in the supernatant of treated cells. The calibrated model was then utilized to determine the number of binding sites of various cell lines. The intracellular model was extended by a pharmacokinetic (PK) and -dynamic (PD) part. By calibrating the combined ESA-EpoR-PKPD model to data from clinical trails, patient-specific parameters could be estimated. In the final step, the combined model was used to retrospectively predict an optimized dosing schedule for all individual patients of the clinical trials. Therefore, the existing treatment of anemia could potentially be optimized by choosing the most adequate ESA and by adapting the dosing schedule on a patient-specific level based on our model predictions.

Many of the model predictions were validated experimentally *in vitro*. However, validation of the optimized treatment strategies would require a whole new clinical trial which exceeded the budged and time frame of the project. Nevertheless, based on our results, a patent [168] could be filed, facilitating the transfer of the obtained knowledge to partners from industry.

For the second part of Chapter 5, we asked:

Does the micro-environment of immune cells affect gene expression? How reliably can the immune cell content in tumor tissue be predicted from blood-derived reference profiles?

Mathematical deconvolution of cell mixtures or bulk tumor samples can be performed based on gene expression data. However, the accurate deconvolution is challenging due to biological and technical noise in the measurements and contamination of the sample caused by cell types for which no reference gene expression profiles exist. To date, existing approaches were mainly based on microarray data and were validated on *in vitro* cell mixtures or whole blood samples [1, 157, 139]. As the gene expression of immune and stromal cells is known to depend on the micro-environment, we performed a benchmark study based on data from three different human tissue sources. Also, modern high-throughput RNA sequencing is less noisy than microarray, especially for genes with low expression numbers. Thus, we used data from single-cell RNA-seq to benchmark the accuracy of existing deconvolution approaches and determined the best working method.

In our data analysis, we found that malignant tumor cells cluster according to the patient while most other cell types cluster by their cell type and tissue source. This indicates that tumor cells might exhibit more patient-specific features as immune and stromal

cells. In our deconvolution benchmark, we could show that reference gene expression profiles (RGEPs) purely based on blood-derived cells are insufficient for deconvolution of bulk tumor samples. Despite the variability of tumor cells from patient to patient, using source tissue specific RGEPs we were able to obtain accurate deconvolution results for most cell types. Furthermore, prognostic ratios of T cell subtypes could be determined making the method valuable for identifying patient subpopulations for immunotherapy treatment. Based on the deconvolution result, we could also predict improved gene expression profiles of malignant tumor cells in a patient-specific manner. This could be of further interest to identify biomarkers and genes that reliably correlate for instance with the abundance of certain other cell types.

This proof of principle study revealed that micro-environment and patient-specific gene expression have an impact on the deconvolution result that is not negligible. In clinical practice, biopsies of malignant tissue and its surroundings are part of the standard procedure. RNA sequencing of these samples is mostly performed as a bulk measurement rather than on the single-cell level due to the reduced costs and effort. For a successful application of the presented approach in a clinical setting, however, further validation on bulk sequencing data would be needed proving the reliable and correct functioning on data obtained from different experimental platforms.

6.2 Outlook

Cancer and infectious disease are among the leading causes of death worldwide. However, several recent developments have the potential to drastically improve the existing prevention measures and treatment strategies. In the following, I want to briefly introduce these approaches and highlight their importance for the progress in fighting diseases. Furthermore, the role of systems biology for the development of novel treatments and for the progress towards precision medicine shall be discussed.

Towards universal influenza vaccines and cell culture-based vaccine production

As mentioned in Section 4.1, the current vaccines against seasonal influenza virus strains rely on predictions for the most prominent strains in circulation. As the strains mutate due to antigenic drift, vaccines require modifications each year [140]. Depending on how well the prediction matches the actual strains in fluctuation, the vaccine effectiveness strongly fluctuates [145]. Another challenge in the development and production of influenza vaccines is the short time frame from strain prediction to large-scale production

for a timely availability at the beginning of the season. Unlike current seasonal vaccines, *universal vaccines* are not specific for an HA or NA antigen sequences but make use of the conserved stalk domain of the HA, NA or M2 proteins [102]. Thereby, these vaccines do not rely on strain predictions and can be used to prevent both, seasonal and pandemic influenza virus infection.

A second factor for improving and accelerating the process of vaccine production relies on cell-based assays rather than growing the viruses in embryonated chicken eggs. For optimizing the virus yield, mathematical models have been developed describing the population dynamics in bioreactors [61] and linking intracellular kinetics to the population scale [75]. This approach could make the vaccine production independent of the supply with embryonated eggs and therefore quickly scale the production to the needs of the market [102].

While the first cell culture-based influenza vaccine was approved recently, for universal vaccines the first pre-clinical studies showed promising results but need further investigation in a clinical setting, as summarized by Krammer & Palese [102].

Novel technologies and their impact cancer treatment – *en route* to precision medicine

The cost for sequencing one human genome sequence dropped from more than $\$\,100\,\mathrm{M}$ in 2001 to less than $\$\,1000$ dollars in 2014 [45]. Therefore, NGS technologies have found their way into clinics and genome-wide sequencing such as RNA-seq is now often part of the clinical routine for patients with cancer. This step enables clinicians to analyze mutations that led to the specific type cancer on a patient-specific level. With this information at hand, not only the therapy can be selected according to the mutation status, for instance based on the HER2 status in breast cancer [94] and the KRAS mutation status in colorectal cancer [155], but also the understanding of disease can be improved in retrospective analyses. Furthermore, novel single-cell technologies, such as scRNA-seq, reveal the heterogeneity of cells within tumor samples [190] and allow for the characterization of immune cells based on their transcriptome-wide gene expression [217].

The use of computational methods for predicting the micro-environment of the tumor for immunotherapy and for optimizing the dosing schedule was extensively discussed in this thesis. A new initiative on precision medicine, as announced by Barack Obama in his *State of the Union Address* in 2015, supports this development by providing additional funding but also by adapting the standards of regulatory authorities, such as the FDA, to allow for the approval of novel treatment strategies [36].

From systems biology to systems medicine

Diseases, such as cancer, involve a multitude of complex signaling pathways and regulatory networks. Therefore, a systems approach is essential to understand the interplay of involved components and eventually to combat the disease. While systems biology has its main focus on basic research, predictive computational models are also used to establish diagnostic biomarkers, identify potential drug targets and to optimize existing treatment strategies. With increasing availability of high-throughput data, models that integrate heterogeneous data from *in vitro* experiments as well as from clinics have become a valuable tool for medical research and diagnostics. Therefore, the transition from systems biology to *systems medicine* is already under way [206].

The work presented in the scope of this thesis is exemplary for this transition. The mechanistic ESA-EpoR multi-scale model presented in Section 5.2 showed that results from "wet lab" can be transferred to predictions for the clinics. Furthermore, the dosing of the drug can be adapted on a patient-specific basis thereby reducing the risk of over- and underdosing and the associated side effects. Also the method presented in Section 5.3 has a clear focus on the medical application. Knowing the cellular composition of tumor samples is key for selecting the appropriate treatment strategy. Using computational deconvolution, this precious information can be obtained from data that is highly available in the clinical context. Furthermore, the prediction of improved tumor gene expression profiles might be used for identifying novel diagnostic biomarkers for immunotherapy. Taken together, predictive computational models promote and accelerate the knowledge transfer from basic research to clinical application.

Bibliography

[1] Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. „Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus“. *PLOS ONE* 2009, **4**(7).

[2] Agarwal, S., Mierle, K., *et al. Ceres Solver*. `http://ceres-solver.org`. 2016.

[3] Amorim, M. J., Bruce, E. a., Read, E. K. C., Foeglein, A., Mahen, R., Stuart, A. D. & Digard, P. „A Rab11- and microtubule-dependent mechanism for cytoplasmic transport of influenza A virus viral RNA.“ *Journal of virology* May 2011, **85**(9): 4143–56.

[4] Arvis, J.-F. & Shepherd, B. *The Air Connectivity Index: Measuring Integration in the Global Air Transport Network*. The World Bank, June 2011.

[5] Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A. & Blom, J. G. „Systems biology: parameter estimation for biochemical models“. *FEBS Journal* Jan. 2009, **276**(4): 886–902.

[6] Avilov, S. V., Moisy, D., Naffakh, N. & Cusack, S. „Influenza A virus progeny vRNP trafficking in live infected cells studied with the virus-encoded fluorescently tagged PB2 protein.“ *Vaccine* Dec. 2012, **30**(51): 7411–7.

[7] Babcock, H. P., Chen, C. & Zhuang, X. „Using single-particle tracking to study nuclear trafficking of viral genes.“ *Biophysical journal* Oct. 2004, **87**(4): 2749–58.

[8] Bachmann, J., Raue, A., Schilling, M., Böhm, M. E., Kreutz, C., Kaschek, D., Busch, H., Gretz, N., Lehmann, W. D., Timmer, J. & Klingmüller, U. „Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range.“ *Molecular Systems Biology* Jan. 2011, **7**(1): 516.

[9] Bancroft, C. T. & Parslow, T. G. „Evidence for segment-nonspecific packaging of the influenza a virus genome.“ *Journal of virology* July 2002, **76**(14): 7133–9.

[10] Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, P., Silva, M. de, Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R. & Garraway, L. A. „The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity“. *Nature* Mar. 2012, **483**(7391): 603–307.

[11] Beauchemin, C. A. A. & Handel, A. „A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead.“ *BMC Public Health* Jan. 2011, **11 Suppl 1**(Suppl 1): S7.

[12] Becker, V., Schilling, M., Bachmann, J., Baumann, U., Raue, A., Maiwald, T., Timmer, J. & Klingmüller, U. „Covering a broad dynamic range: information processing at the erythropoietin receptor.“ *Science* June 2010, **328**(5984): 1404–8.

[13] Bellavia, S., Macconi, M. & Morini, B. „STRSCNE: A Scaled Trust-Region Solver for Constrained Nonlinear Equations“. *Computational Optimization and Applications* Apr. 2004, **28**(1): 31–50.

[14] Bevan, A. *Statistical Data Analysis for the Physical Sciences*. Cambridge: Cambridge University Press, 2013.

[15] Binder, M., Sulaimanov, N., Clausznitzer, D., Schulze, M., Hüber, C. M., Lenz, S. M., Schlöder, J. P., Trippler, M., Bartenschlager, R., Lohmann, V. & Kaderali, L. „Replication Vesicles are Load- and Choke-Points in the Hepatitis C Virus Lifecycle“. *PLoS Pathogens* Aug. 2013, **9**(8): 1–21.

[16] Birgegård, G., Aapro, M. S., Bokemeyer, C., Dicato, M., Drings, P., Hornedo, J., Krzakowski, M., Ludwig, H., Pecorelli, S., Schmoll, H.-J., Schneider, M., Schrijvers, D., Shasha, D. & Biesen, S. van. „Cancer-Related Anemia: Pathogenesis, Prevalence and Treatment“. *Oncology* Apr. 2005, **68**(1): 3–11.

[17] Bornstein, B. J., Keating, S. M., Jouraku, A. & Hucka, M. „LibSBML: an API Library for SBML“. *Bioinformatics* Feb. 2008, **24**(6): 880–881.

[18] Bouvier, N. M. & Palese, P. „The biology of influenza viruses“. *Vaccine* Sept. 2008, **26**: D49–D53.

[19] Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. „Near-optimal probabilistic RNA-seq quantification“. *Nature Biotechnology* Apr. 2016, **34**(5): 525–527.

[20] Brockmann, D. & Helbing, D. „The Hidden Geometry of Complex, Network-Driven Contagion Phenomena“. *Science* Dec. 2013, **342**(6164): 1337–1342.

[21] Broudy, V., Lin, N., Brice, M., Nakamoto, B. & Papayannopoulou, T. „Erythropoietin receptor characteristics on primary human erythroid cells“. *Blood* 1991, **77**(12): 2583–2590.

[22] Butcher, J. *Numerical Methods for Ordinary Differential Equations*. New York: Wiley, 2008.

[23] Chan, M. *Influenza A(H1N1) – Statement by WHO Director-General, Dr Margaret Chan*. `http://www.who.int/mediacentre/news/statements/2009/h1n1_20090429/en/`. [Online; accessed 19-April-2017]. 2009.

[24] Chang, C.-C. & Lin, C.-J. „LIBSVM: A Library for Support Vector Machines“. *ACM Transactions on Intelligent Systems and Technology* 2011, **2**: 27:1–27:27.

[25] Chang, C.-C. & Lin, C.-J. „Training v-Support Vector Regression: Theory and Algorithms“. *Neural Computation* Aug. 2002, **14**(8): 1959–1977.

[26] Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. *shiny: Web Application Framework for R*. R package version 0.11.1. 2015.

[27] Chanu, P., Mu, S. & Reigner, B. Research Report No. 1026722 (Clinical Study Report NH19960 / NCT00327535). (F. Hoffmann La Roche, Ltd, 2007). Clinical trial. 2007.

[28] Chen, D. S. & Mellman, I. „Oncology meets immunology: The cancer-immunity cycle". *Immunity* 2013, **39**(1): 1–10.

[29] Chen, W., Hepburn, I. & De Schutter, E. „Tetrahedral mesh generation and visualization for stochastic reaction-diffusion simulation". *BMC Neuroscience* 2010, **11**(Suppl 1): P70.

[30] Cheung, T. K. W. & Poon, L. L. M. „Biology of influenza A virus." *Annals of the New York Academy of Sciences* Apr. 2007, **1102**: 1–25.

[31] Chou, Y.-y., Heaton, N. S., Gao, Q., Palese, P., Singer, R. & Lionnet, T. „Colocalization of different influenza viral RNA segments in the cytoplasm before viral budding as shown by single-molecule sensitivity FISH analysis." *PLoS Pathogens* May 2013, **9**(5): e1003358.

[32] Chou, Y.-y., Vafabakhsh, R., Doğanay, S., Gao, Q., Ha, T. & Palese, P. „One influenza virus particle packages eight unique viral RNAs as shown by FISH analysis." *Proceedings of the National Academy of Sciences of the United States of America* June 2012, **109**(23): 9101–6.

[33] Clements, B. W. & Casani, J. A. P. „Pandemic Influenza". In: Disasters and Public Health. Elsevier, 2016: 385–410.

[34] Coleman, T. F. & Li, Y. „An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds". *SIAM Journal on Optimization* 1996, **6**: 418–445.

[35] Coleman, T. F. & Li, Y. „On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds". *Mathematical Programming* 1994, **67**(1-3): 189–224.

[36] Collins, F. S. & Varmus, H. „A New Initiative on Precision Medicine". *New England Journal of Medicine* Feb. 2015, **372**(9): 793–795.

[37] Cros, J. F. & Palese, P. „Trafficking of viral genomic RNA into and out of the nucleus: influenza, Thogoto and Borna disease viruses". *Virus Research* Sept. 2003, **95**(1-2): 3–12.

[38] Davies, B. E. „Pharmacokinetics of oseltamivir: an oral antiviral for the treatment and prophylaxis of influenza in diverse populations." *The Journal of antimicrobial chemotherapy* 2010, **65 Suppl 2**: 5–10.

[39] Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P.-Y., Bandaranayake, D., Breiman, R. F., Brooks, W. A., Buchy, P., Feikin, D. R., Fowler, K. B., Gordon, A., Hien, N. T., Horby, P., Huang, Q. S., Katz, M. A., Krishnan, A., Lal, R., Montgomery, J. M., Mølbak, K., Pebody, R., Presanis, A. M., Razuri, H., Steens, A., Tinoco, Y. O., Wallinga, J., Yu, H., Vong, S., Bresee, J. & Widdowson, M.-A. „Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study". *The Lancet Infectious Diseases* Sept. 2012, **12**(9): 687–695.

[40] Dee, K. U. & Shuler, M. L. „A mathematical model of the trafficking of acid-dependent enveloped viruses: application to the binding, uptake, and nuclear accumulation of baculovirus." *Biotechnology and bioengineering* June 1997, **54**(5): 468–90.

[41] Dee, K. U., Hammer, D. a. & Shuler, M. L. „A model of the binding, entry, uncoating, and RNA synthesis of Semliki Forest virus in baby hamster kidney (BHK-21) cells.“ *Biotechnology and bioengineering* 1995, **46**: 485–496.

[42] DeLuca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W. & Getz, G. „RNA-SeQC: RNA-seq metrics for quality control and process optimization“. *Bioinformatics* Apr. 2012, **28**(11): 1530–1532.

[43] Dennis, J. E., Gay, D. M. & Welsch, R. E. „Algorithm 573: NL2SOL—An Adaptive Nonlinear Least-Squares Algorithm [E4]“. *ACM Transactions on Mathematical Software* Sept. 1981, **7**(3): 369–383.

[44] Dickinson, R. P. & Gelinas, R. J. „Sensitivity analysis of ordinary differential equation systems—A direct method“. *Journal of Computational Physics* June 1976, **21**(2): 123–143.

[45] Dijk, E. L. van, Auger, H., Jaszczyszyn, Y. & Thermes, C. „Ten years of next-generation sequencing technology“. *Trends in Genetics* Sept. 2014, **30**(9): 418–426.

[46] Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A. & Wang, W. „Normalization and noise reduction for single cell RNA-seq experiments“. *Bioinformatics* Feb. 2015, **31**(13): 2225–2227.

[47] Dobin, A., Davis, C. a., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. „STAR: Ultrafast universal RNA-seq aligner“. *Bioinformatics* 2013, **29**(1): 15–21.

[48] DuBois, R. M., Zaraket, H., Reddivari, M., Heath, R. J., White, S. W. & Russell, C. J. „Acid stability of the hemagglutinin protein regulates H5N1 influenza virus pathogenicity.“ *PLoS Pathogens* Dec. 2011, **7**(12): e1002398.

[49] Dunlop, E., Percy, M., Boland, M., Maxwell, A. & Lappin, T. „Induction of Signalling in Non-Erythroid Cells by Pharmacological Levels of Erythropoietin“. *Neurodegenerative Diseases* May 2006, **3**(1-2): 94–100.

[50] Efron, B. & Stein, C. „The Jackknife Estimate of Variance“. *The Annals of Statistics* May 1981, **9**(3): 586–596.

[51] Egrie, J. C. & Browne, J. K. „Development and characterization of novel erythropoiesis stimulating protein (NESP)“. *Nephrology Dialysis Transplantation* June 2001, **16**(suppl 3): 3–13.

[52] Eisenberg, E. & Levanon, E. Y. „Human housekeeping genes, revisited“. *Trends in Genetics* Oct. 2013, **29**(10): 569–574.

[53] Elliott, S. „Bad Science: Cause and Consequence“. *Journal of Pharmaceutical Sciences* Apr. 2016, **105**(4): 1358–1361.

[54] Elliott, S., Egrie, J., Browne, J., Lorenzini, T., Busse, L., Rogers, N. & Ponting, I. „Control of rHuEPO biological activity: The role of carbohydrate“. *Experimental Hematology* Dec. 2004, **32**(12): 1146–1155.

[55] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. „Stochastic gene expression in a single cell.“ *Science* Aug. 2002, **297**(5584): 1183–6.

[56] Engvall, E. & Perlmann, P. „Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G“. *Immunochemistry* Sept. 1971, **8**(9): 871–874.

[57] Enzoklop. *Polymerase chain reaction*. `https://commons.wikimedia.org/wiki/File:Polymerase_chain_reaction.svg` under CC BY-SA 3.0 (`https://creativecommons.org/licenses/by-sa/3.0/legalcode`). [Online; accessed 28-April-2017]. 2014.

[58] Ester, M., Kriegel, H. P., Sander, J. & Xu, X. „A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise“. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. 1996: 226–231.

[59] Fange, D., Mahmutovic, A. & Elf, J. „MesoRD 1.0: Stochastic reaction-diffusion simulations in the microscopic limit“. *Bioinformatics* Oct. 2012, **28**(23): 3155–3157.

[60] Forsberg, E. C., Serwold, T., Kogan, S., Weissman, I. L. & Passegué, E. „New Evidence Supporting Megakaryocyte-Erythrocyte Potential of Flk2/Flt3+ Multipotent Hematopoietic Progenitors“. *Cell* July 2006, **126**(2): 415–426.

[61] Frensing, T., Heldt, F. S., Pflugmacher, A., Behrendt, I., Jordan, I., Flockerzi, D., Genzel, Y. & Reichl, U. „Continuous influenza virus production in cell culture shows a periodic accumulation of defective interfering particles.“ *PLOS ONE* Jan. 2013, **8**(9): e72288.

[62] Fulwyler, M. J. „Electronic Separation of Biological Cells by Volume“. *Science* Nov. 1965, **150**(3698): 910–911.

[63] Galloway, S. E., Reed, M. L., Russell, C. J. & Steinhauer, D. A. „Influenza HA Subtypes Demonstrate Divergent Phenotypes for Cleavage Activation and pH of Fusion: Implications for Host Range and Adaptation“. *PLoS Pathogens* Feb. 2013, **9**(2). Ed. by Kawaoka, Y.: e1003151.

[64] Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoué, F., Bruneval, P., Cugnenc, P.-H., Trajanoski, Z., Fridman, W.-H. & Pagès, F. „Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome“. *Science* 2006, **313**(5795): 1960–1964.

[65] Gardiner, C. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. 4th. Berlin, Heidelberg: Springer, 2009: 447.

[66] Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. „Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins“. *Molecular & Cellular Proteomics* Jan. 2012, **11**(3): M111.014050–M111.014050.

[67] Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., Nair, V. S., Xu, Y., Khuong, A., Hoang, C. D., Diehn, M., West, R. B., Plevritis, S. K. & Alizadeh, A. a. „The prognostic landscape of genes and infiltrating immune cells across human cancers“. *Nature Medicine* 2015, (July).

[68] Gillespie, D. T. „Exact stochastic simulation of coupled chemical reactions“. *The Journal of Physical Chemistry* Dec. 1977, **81**(25): 2340–2361.

[69] Gong, T. & Szustakowski, J. D. „DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data“. *Bioinformatics* Feb. 2013, **29**(8): 1083–1085.

[70] Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S. & Szustakowski, J. D. „Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples“. *PLOS ONE* Nov. 2011, **6**(11). Ed. by Rattray, M.: e27156.

[71] Goodwin, S., McPherson, J. D. & McCombie, W. R. „Coming of age: ten years of next-generation sequencing technologies“. *Nature Reviews Genetics* May 2016, **17**(6): 333–351.

[72] Grant, M. & Boyd, S. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*. `http://cvxr.com/cvx`. Mar. 2014.

[73] Gross, A. W. & Lodish, H. F. „Cellular Trafficking and Degradation of Erythropoietin and Novel Erythropoiesis Stimulating Protein (NESP)“. *Journal of Biological Chemistry* Nov. 2005, **281**(4): 2024–2032.

[74] Heinrich, R. & Schuster, S. *The Regulation of Cellular Systems*. Springer Science + Business Media, 1996.

[75] Heldt, F. S., Frensing, T., Pflugmacher, A., Gröpler, R., Peschel, B. & Reichl, U. „Multiscale Modeling of Influenza A Virus Infection Supports the Development of Direct-Acting Antivirals“. *PLoS Computational Biology* Nov. 2013, **9**(11). Ed. by Koelle, K.: e1003372.

[76] Heldt, F. S., Frensing, T. & Reichl, U. „Modeling the intracellular dynamics of influenza virus replication to understand the control of viral RNA synthesis.“ *Journal of virology* Aug. 2012, **86**(15): 7806–17.

[77] Hengl, S., Kreutz, C., Timmer, J. & Maiwald, T. „Data-based identifiability analysis of non-linear dynamical models“. *Bioinformatics* July 2007, **23**(19): 2612–2618.

[78] Henke, M., Mattern, D., Pepe, M., Bézay, C., Weissenberger, C., Werner, M. & Pajonk, F. „Do Erythropoietin Receptors on Cancer Cells Explain Unexpected Clinical Findings?“ *Journal of Clinical Oncology* Oct. 2006, **24**(29): 4708–4713.

[79] Hepburn, I., Chen, W., Wils, S. & De Schutter, E. „STEPS: efficient simulation of stochastic reaction-diffusion models in realistic morphologies.“ *BMC Systems Biology* May 2012, **6**(1): 36.

[80] Herzenberg, L. A., Parks, D., Sahaf, B., Perez, O., Roederer, M. & Herzenberg, L. A. „The History and Future of the Fluorescence Activated Cell Sorter and Flow Cytometry: A View from Stanford“. *Clinical Chemistry 48:10* 2002, **48**(10): 1819–1827.

[81] Hindmarsh, A. C., Brown, P. N., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E. & Woodward, C. S. „SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers“. *ACM Trans. Math. Softw.* Sept. 2005, **31**(3): 363–396.

[82] Hirsh, V., Eid, J. & Yoo, K. Research Report No. 1020102 (Clinical Study Report NA17101 / NCT0072059). (Royal Victoria Hospital Montreal QC H3A 1A1 Canada and Hoffmann-La Roche Inc, NJ USA, 2006). Clinical trial. 2006.

[83] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. & Kummer, U. „COPASI–a COmplex PAthway SImulator“. *Bioinformatics* Oct. 2006, **22**(24): 3067–3074.

[84] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. & Wang, J. „The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models". *Bioinformatics* 2003, **19**(4): 524–531.

[85] Huotari, J. & Helenius, A. „Endosome maturation." *EMBO Journal* 2011, **30**: 3481–3500.

[86] Hutchinson, E. C., Kirchbach, J. C. von, Gog, J. R. & Digard, P. „Genome packaging in influenza A virus." *The Journal of general virology* Feb. 2010, **91**(Pt 2): 313–28.

[87] Imai, K. & Takaoka, A. „Comparing antibody and small-molecule therapies for cancer". *Nature Reviews Cancer* Sept. 2006, **6**(9): 714–727.

[88] Ingalls, B. *Mathematical Modeling in Systems Biology: An Introduction*. Cambridge: MIT Press, 2013.

[89] Ingalls, B. P. & Sauro, H. M. „Sensitivity analysis of stoichiometric networks: an extension of metabolic control analysis to non-steady state trajectories". *Journal of Theoretical Biology* May 2003, **222**(1): 23–36.

[90] Ino, Y., Yamazaki-Itoh, R., Shimada, K., Iwasaki, M., Kosuge, T., Kanai, Y. & Hiraoka, N. „Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer". *British Journal of Cancer* Feb. 2013, **108**(4): 914–923.

[91] Islam, S., Zeisel, A., Joost, S., Manno, G. L., Zajac, P., Kasper, M., Lönnerberg, P. & Linnarsson, S. „Quantitative single-cell RNA-seq with unique molecular identifiers". *Nature Methods* Dec. 2013, **11**(2): 163–166.

[92] Iwasaki, A. & Pillai, P. S. „Innate immunity to influenza virus infection". *Nature Reviews Immunology* Apr. 2014, **14**(5): 315–328.

[93] Jarsch, M., Brandt, M., Lanzendörfer, M. & Haselbeck, A. „Comparative Erythropoietin Receptor Binding Kinetics of C.E.R.A. and Epoetin-β Determined by Surface Plasmon Resonance and Competition Binding Assay". *Pharmacology* Sept. 2007, **81**(1): 63–69.

[94] Kamalakaran, S., Lezon-Geyda, K., Varadan, V., Banerjee, N., Lannin, D. R., Rizack, T., Sikov, W. M., Abu-Khalaf, M. M., Janevski, A. & and, L. H. „Evaluation of ER/PR and HER2 status by RNA sequencing in tissue core biopsies from preoperative clinical trial specimens." *Journal of Clinical Oncology* Sept. 2011, **29**(27_suppl): 46–46.

[95] Karlas, A., Machuy, N., Shin, Y., Pleissner, K.-P., Artarini, A., Heuer, D., Becker, D., Khalil, H., Ogilvie, L. A., Hess, S., Mäurer, A. P., Müller, E., Wolff, T., Rudel, T. & Meyer, T. F. „Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication". *Nature* Jan. 2010, **463**(7282): 818–822.

[96] Keating, S. M., Smith, L. P., Shapiro, B. E., Hucka, M., Bergmann, F. T., Olivier, B. & Finney, A. *SBML Test Suite Semantic Test Cases 3.2.0*. 2016.

[97] Kirkpatrick, C., B. Reigner, B. & Jordan, P. Research Report No. 1018354 (Clinical Study Report BP18035). (Roche Products Ltd. and Hoffmann-La Roche Ltd., 2005). Clinical trial. 2005.

[98] Kitano, H. „Computational systems biology“. *Nature* Nov. 2002, **420**(6912): 206–210.

[99] Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. & Taipale, J. „Counting absolute numbers of molecules using unique molecular identifiers“. *Nature Methods* Nov. 2011, **9**(1): 72–74.

[100] Klipp, E., Wade, R. C. & Kummer, U. „Biochemical network-based drug-target prediction.“ *Current Opinion in Biotechnology* Aug. 2010, **21**(4): 511–6.

[101] König, R., Stertz, S., Zhou, Y., Inoue, A., Hoffmann, H.-H., Bhattacharyya, S., Alamares, J. G., Tscherne, D. M., Ortigoza, M. B., Liang, Y., Gao, Q., Andrews, S. E., Bandyopadhyay, S., De Jesus, P., Tu, B. P., Pache, L., Shih, C., Orth, A., Bonamy, G., Miraglia, L., Ideker, T., García-Sastre, A., Young, J. a. T., Palese, P., Shaw, M. L. & Chanda, S. K. „Human host factors required for influenza virus replication.“ *Nature* Feb. 2010, **463**(7282): 813–7.

[102] Krammer, F. & Palese, P. „Advances in the development of influenza virus vaccines“. *Nature Reviews Drug Discovery* Feb. 2015, **14**(3): 167–182.

[103] Kreijtz, J., Fouchier, R. & Rimmelzwaan, G. „Immune responses to influenza virus infection“. *Virus Research* Dec. 2011, **162**(1-2): 19–30.

[104] Kreutz, C., Rodriguez, M. B., Maiwald, T., Seidl, M., Blum, H., Mohr, L. & Timmer, J. „An error model for protein quantification“. *Bioinformatics* Sept. 2007, **23**(20): 2747–2753.

[105] Kreutz, C., Raue, A., Kaschek, D. & Timmer, J. *FEBS Journal* June 2013, **280**(11): 2564–71.

[106] Kreutz, C., Raue, A. & Timmer, J. „Likelihood based observability analysis and confidence intervals for predictions of dynamic models“. *BMC Systems Biology* 2012, **6**(1): 120.

[107] Kublun, I., Ehm, P., Brehm, M. A. & Nalaskowski, M. M. „Efficacious inhibition of Importin α/β-mediated nuclear import of human inositol phosphate multikinase“. *Biochimie* 2014, **102**(1): 117–123.

[108] Lagache, T. & Holcman, D. „Effective Motion of a Virus Trafficking Inside a Biological Cell“. *SIAM Journal on Applied Mathematics* Jan. 2008, **68**(4): 1146–1167.

[109] Lake, R. A. & Robinson, B. W. „Opinion: Immunotherapy and chemotherapy — a practical partnership“. *Nature Reviews Cancer* Apr. 2005, **5**(5): 397–405.

[110] Lander, E. S., Linton, L. M., Birren, B., *et al.* „Initial sequencing and analysis of the human genome“. *Nature* Feb. 2001, **409**(6822): 860–921.

[111] Langmead, B. & Salzberg, S. L. „Fast gapped-read alignment with Bowtie 2“. *Nature Methods* Mar. 2012, **9**(4): 357–359.

[112] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. „Ultrafast and memory-efficient alignment of short DNA sequences to the human genome“. *Genome Biology* 2009, **10**(3): R25.

[113] Larson, E., Dominik, J., Rowberg, A. & Higbee, G. „Influenza virus population dynamics in the respiratory tract of experimentally infected mice“. *Infect. Immun.* 1976, **13**(2): 438–447.

[114] Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. „Gradient-based learning applied to document recognition“. *Proceedings of the IEEE* 1998, **86**(11): 2278–2324.

[115] Li, B. & Dewey, C. N. „RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.“ *BMC Bioinformatics* 2011, **12**: 323.

[116] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. a. & Dewey, C. N. „RNA-Seq gene expression estimation with read mapping uncertainty“. *Bioinformatics* 2009, **26**(4): 493–500.

[117] Li, P., Piao, Y., Shon, H. S. & Ryu, K. H. „Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data“. *BMC Bioinformatics* 2015, **16**(1): 347.

[118] Lichtman, J. W. & Conchello, J.-A. „Fluorescence microscopy“. *Nature Methods* Dec. 2005, **2**(12): 910–919.

[119] MacDonald, N. „Time delay in simple chemostat models“. *Biotechnology and Bioengineering* 1976, **18**(6): 805–812.

[120] Macdougall, I. C. „Novel Erythropoiesis-Stimulating Agents: A New Era in Anemia Management“. *Clinical Journal of the American Society of Nephrology* Jan. 2008, **3**(1): 200–207.

[121] Macdougall, I. C. „CERA (Continuous Erythropoietin Receptor Activator): a new erythropoiesis-stimulating agent for the treatment of anemia.“ *Current hematology reports* Nov. 2005, **4**(6): 436–40.

[122] Madrahimov, A., Helikar, T., Kowal, B., Lu, G. & Rogers, J. „Dynamics of Influenza Virus and Human Host Interactions During Infection and Replication Cycle.“ *Bulletin of mathematical biology* Oct. 2012.

[123] Mailankody, S. & Prasad, V. „Five Years of Cancer Drug Approvals“. *JAMA Oncology* July 2015, **1**(4): 539.

[124] Maiwald, T. & Timmer, J. „Dynamical modeling and multi-experiment fitting with PottersWheel.“ *Bioinformatics* Sept. 2008, **24**(18): 2037–43.

[125] Mardis, E. R. „Next-Generation DNA Sequencing Methods“. *Annual Review of Genomics and Human Genetics* Sept. 2008, **9**(1): 387–402.

[126] Martin, J. A. & Wang, Z. „Next-generation transcriptome assembly“. *Nature Reviews Genetics* Sept. 2011, **12**(10): 671–682.

[127] Martin, K. & Helenius, a. „Nuclear transport of influenza virus ribonucleoproteins: the viral matrix protein (M1) promotes export and inhibits import.“ *Cell* Oct. 1991, **67**(1): 117–30.

[128] Matsuoka, Y., Matsumae, H., Katoh, M., Eisfeld, A. J., Neumann, G., Hase, T., Ghosh, S., Shoemaker, J. E., Lopes, T. J., Watanabe, T., Watanabe, S., Fukuyama, S., Kitano, H. & Kawaoka, Y. „A comprehensive map of the influenza A virus replication cycle“. *BMC Systems Biology* 2013, **7**(1): 97.

[129] McClellan, W., Aronoff, S. L., Bolton, W. K., Hood, S., Lorber, D. L., Tang, K. L., Tse, T. F., Wasserman, B. & Leiserowitz, M. „The prevalence of anemia in patients with chronic kidney disease“. *Current Medical Research and Opinion* Aug. 2004, **20**(9): 1501–1510.

[130] Mehnert, J. M., Panda, A., Zhong, H., Hirshfield, K., Damare, S., Lane, K., Sokol, L., Stein, M. N., Rodriguez-Rodriquez, L., Kaufman, H. L., Ali, S., Ross, J. S., Pavlick, D. C., Bhanot, G., White, E. P., DiPaola, R. S., Lovell, A., Cheng, J. & Ganesan, S. „Immune activation and response to pembrolizumab in POLE-mutant endometrial cancer“. *Journal of Clinical Investigation* May 2016, **126**(6): 2334–2340.

[131] Merkle, R., Steiert, B., Salopiata, F., Depner, S., Raue, A., Iwamoto, N., **Schelker, M.**, Hass, H., Wäsch, M., Böhm, M. E., Mücke, O., Lipka, D. B., Plass, C., Lehmann, W. D., Kreutz, C., Timmer, J., Schilling, M. & Klingmüller, U. „Identification of Cell Type-Specific Differences in Erythropoietin Receptor Signaling in Primary Erythroid and Lung Cancer Cells“. *PLOS Computational Biology* Aug. 2016, **12**(8). Ed. by Mac Gabhann, F.: e1005049.

[132] Metzker, M. L. „Sequencing technologies — the next generation“. *Nature Reviews Genetics* Dec. 2009, **11**(1): 31–46.

[133] Michaelis, L. & Menten, M. I. „Die Kinetik der Invertinwirkung“. *Biochemische Zeitschrift* 1913, **49**: 333–369.

[134] Mohammadi, S., Zuckerman, N., Goldsmith, A. & Grama, A. „A Critical Survey of Deconvolution Methods for Separating cell-types in Complex Tissues“. *ArXiv e-prints* Oct. 2015.

[135] Momose, F., Sekimoto, T., Ohkura, T., Jo, S., Kawaguchi, A., Nagata, K. & Morikawa, Y. „Apical transport of influenza A virus ribonucleoprotein requires Rab11-positive recycling endosome.“ *PLOS ONE* Jan. 2011, **6**(6): e21123.

[136] Morini, B. & Porcelli, M. „TRESNEI, a Matlab trust-region solver for systems of nonlinear equalities and inequalities“. *Computational Optimization and Applications* Apr. 2010, **51**(1): 27–49.

[137] Mukherjee, S. *An Oncologist's Pulitzer-Winning Cancer Biography*. `http://www.npr.org/2011/04/22/135600761/an-oncologists-pulitzer-winning-cancer-biography?ft=3&f=1003,1004,1007,1013,1014,1017,1019,1128`. [Online; accessed 19-April-2017]. 2011.

[138] Newman, A. M. & Alizadeh, A. A. „High-throughput genomic profiling of tumor-infiltrating leukocytes“. *Current Opinion in Immunology* 2016, **41**: 77–84.

[139] Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M. & Alizadeh, A. A. „Robust enumeration of cell subsets from tissue expression profiles“. *Nature Methods* Mar. 2015, **12**(5): 453–457.

[140] Nichol, K. & Treanor, J. „Vaccines for Seasonal and Pandemic Influenza“. *The Journal of Infectious Diseases* Nov. 2006, **194**(s2): S111–S118.

[141] Nieba, L., Krebber, A. & Plückthun, A. „Competition BIAcore for Measuring True Affinities: Large Differences from Values Determined from Binding Kinetics“. *Analytical Biochemistry* Feb. 1996, **234**(2): 155–165.

[142] Noble, D. *The Music of Life: Biology Beyond Genes*. Popular science. Oxford: Oxford University Press, 2008.

[143] Nocedal, J. & Wright, S. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. New York: Springer, 2006.

[144] Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N. & Marathe, M. V. „A systematic review of studies on forecasting the dynamics of influenza outbreaks“. *Influenza and other Respiratory Viruses* 2014, **8**(3): 309–316.

[145] Osterholm, M. T., Kelley, N. S., Sommer, A. & Belongia, E. A. „Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis“. *The Lancet Infectious Diseases* Jan. 2012, **12**(1): 36–44.

[146] Owen, A. A. B. „A Central Limit Theorem for Latin Hypercube Sampling“. *Journal of the Royal Statistical Society. Series B (Methodological)* 1992, **54**(2): 541–551.

[147] Pagès, F., Galon, J., Dieu-Nosjean, M.-C., Tartour, E., Sautès-Fridman, C. & Fridman, W.-H. „Immune infiltration in human tumors: a prognostic factor that should not be ignored“. *Oncogene* Nov. 2009, **29**(8): 1093–1102.

[148] Palese, P. „Influenza: old and new threats“. *Nature Medicine* Dec. 2004, **10**(12s): S82–S87.

[149] Pardoll, D. M. „The blockade of immune checkpoints in cancer immunotherapy“. *Nature Reviews Cancer* 2012, **12**(4): 252–264.

[150] Patro, R., Duggal, G. & Kingsford, C. *Accurate, fast, and model-aware transcript expression quantification with Salmon*. Tech. rep. June 2015.

[151] Patterson, K. D. & Pyle, G. F. „The geography and mortality of the 1918 influenza pandemic.“ *Bulletin of the history of medicine* 1991, **65**(1): 4–21.

[152] Pearson, K. „LIII.On lines and planes of closest fit to systems of points in space“. *Philosophical Magazine Series 6* Nov. 1901, **2**(11): 559–572.

[153] Perfetto, S. P., Chattopadhyay, P. K. & Roederer, M. „Innovation: Seventeen-colour flow cytometry: unravelling the immune system“. *Nature Reviews Immunology* Aug. 2004, **4**(8): 648–655.

[154] Pezze, P. D., Sonntag, A. G., Thien, A., Prentzell, M. T., Godel, M., Fischer, S., Neumann-Haefelin, E., Huber, T. B., Baumeister, R., Shanley, D. P. & Thedieck, K. „A Dynamic Network Model of mTOR Signaling Reveals TSC-Independent mTORC2 Regulation“. *Science Signaling* Mar. 2012, **5**(217): ra25–ra25.

[155] Phipps, A. I., Buchanan, D. D., Makar, K. W., Win, A. K., Baron, J. A., Lindor, N. M., Potter, J. D. & Newcomb, P. A. „KRAS-mutation status in relation to colorectal cancer survival: the joint impact of correlated tumour markers“. *British Journal of Cancer* Mar. 2013, **108**(8): 1757–1764.

[156] Pichlmair, A., Schulz, O., Tan, C. P., Naslund, T. I., Liljestrom, P., Weber, F. & Sousa, C. R. e. „RIG-I-Mediated Antiviral Responses to Single-Stranded RNA Bearing 5'-Phosphates". *Science* Nov. 2006, **314**(5801): 997–1001.

[157] Qiao, W., Quon, G., Csaszar, E., Yu, M., Morris, Q. & Zandstra, P. W. „PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions". *PLoS Comput Biol* Dec. 2012, **8**(12). Ed. by Bonneau, R.: e1002838.

[158] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015.

[159] Raue, A., Karlsson, J., Saccomani, M. P., Jirstrand, M. & Timmer, J. „Comparison of approaches for parameter identifiability analysis of biological systems". *Bioinformatics* 2014, **30**(10): 1440–1448.

[160] Raue, A., Steiert, B., **Schelker, M.**, Kreutz, C., Maiwald, T., Hass, H., Vanlier, J., Tönsing, C., Adlung, L., Engesser, R., Mader, W., Heinemann, T., Hasenauer, J., Schilling, M., Höfer, T., Klipp, E., Theis, F., Klingmüller, U., Schöberl, B. & Timmer, J. „Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems." *Bioinformatics* Nov. 2015, **31**(21): 3558–3560.

[161] Raue, A. *arNLS*. `http://data2dynamics.org`. 2014.

[162] Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. & Timmer, J. *Bioinformatics* Aug. 2009, **25**(15): 1923–9.

[163] Raue, A., Kreutz, C., Theis, F. J. & Timmer, J. „Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability." *Philos. Trans. A. Math. Phys. Eng. Sci.* Mar. 2013, **371**(1984): 20110544.

[164] Raue, A., Schilling, M., Bachmann, J., Matteson, A., **Schelker, M.**, Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U. & Timmer, J. „Lessons Learned from Quantitative Dynamical Modeling in Systems Biology". *PLOS ONE* 2013, **8**(9): e74335.

[165] Reed, M. L., Bridges, O. A., Seiler, P., Kim, J.-K., Yen, H.-L., Salomon, R., Govorkova, E. A., Webster, R. G. & Russell, C. J. „The pH of activation of the hemagglutinin protein regulates H5N1 influenza virus pathogenicity and transmissibility in ducks." *Journal of Virology* 2010, **84**: 1527–1535.

[166] Richards, J., Reigner, B. & Jordan, P. Research Report No. 1007694 (Clinical Study Report WP16422). (PPD Development Clinic, 72 Hospital Close, Evington, Leicester, England and Hoffmann-La Roche Inc Nutley, NJ USA., 2002). Clinical trial. 2002.

[167] Robertson, J. S., Schubert, M. & Lazzarini, R. A. „Polyadenylation sites for influenza virus mRNA." *Journal of virology* 1981, **38**(1): 157–63.

[168] Rodriguez, A., Schilling, M., Klingmüller, U., Raue, A., **Schelker, M.**, Timmer, J., Jarsch, M. & Steiert, B. „Methods for the prediction of a personalized ESA-dose in the treatment of anemia". *WO Patent App. PCT/EP2015/063,775* 2016.

[169] Rodriguez-Fernandez, M., Egea, J. A. & Banga, J. R. „Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems". *BMC Bioinformatics* 2006, **7**(1): 483.

[170] Rodriguez-Gonzalez, A., **Schelker, M.**, Raue, A., Steiert, B., Böhm, M., Salopiata, F., Adlung, L., Stepath, M., Depner, S., Wagner, M.-C., Merkle, R., Kramer, B. A., Lattermann, S., Wäsch, M., Franke, A., Klipp, E., Wuchter, P., Ho, A. D., Lehmann, W. D., Jarsch, M., Schilling, M., Timmer, J. & Klingmüller, U. „Mechanistic multiscale modelling enables personalized treatment". *Manuscript in preparation* 2017.

[171] Ruiz, P., Balcke, P., Martinez, J. M. & Harris, K. „Tolerability of the Epoetin-Beta Multidose Formulation (Reco-Pen®) in Patients with Renal Anaemia". *Clinical Drug Investigation* Sept. 2000, **20**(3): 151–158.

[172] Sadewasser, A., Paki, K., Eichelbaum, K., Bogdanow, B., Saenger, S., Budt, M., Lesch, M., Hinz, K.-P., Herrmann, A., Meyer, T. F., Karlas, A., Selbach, M. & Wolff, T. „Quantitative proteomic approach identifies Vpr binding protein as novel host factor supporting influenza A virus infections in human cells". *Molecular & Cellular Proteomics* Mar. 2017: mcp.M116.065904.

[173] Samji, T. „Influenza A: understanding the viral life cycle." *The Yale journal of biology and medicine* Dec. 2009, **82**(4): 153–9.

[174] Sankaran, V. G. & Weiss, M. J. „Anemia: progress in molecular mechanisms and therapies". *Nature Medicine* Mar. 2015, **21**(3): 221–230.

[175] **Schelker, M.**, Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B. & Raue, A. „Estimation of immune cell content in tumour tissue using single-cell RNA-seq data". *Manuscript in preparation* 2017.

[176] **Schelker, M.**, Mair, C. M., Jolmes, F., Welke, R.-W., Klipp, E., Herrmann, A., Flöttmann, M. & Sieben, C. „Viral RNA Degradation and Diffusion Act as a Bottleneck for the Influenza A Virus Infection Efficiency". *PLOS Computational Biology* 2016, **12**(10): e1005075.

[177] Schilling, M., Maiwald, T., Bohl, S., Kollmann, M., Kreutz, C., Timmer, J. & Klingmüller, U. „Computational processing and error reduction strategies for standardized quantitative data in biological networks". *FEBS Journal* 2005, **272**: 6400–6411.

[178] Schmidt, H. & Jirstrand, M. „Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology." *Bioinformatics* Feb. 2006, **22**(4): 514–5.

[179] Schuss, Z., Singer, A. & Holcman, D. „The narrow escape problem for diffusion in cellular microdomains". *Proceedings of the National Academy of Sciences* Sept. 2007, **104**(41): 16098–16103.

[180] Segel, L. A. & Slemrod, M. „The Quasi-Steady-State Assumption: A Case Study in Perturbation". *SIAM Review* Sept. 1989, **31**(3): 446–477.

[181] Shapira, S. D., Gat-Viks, I., Shum, B. O. V., Dricot, A., Grace, M. M. de, Wu, L., Gupta, P. B., Hao, T., Silver, S. J., Root, D. E., Hill, D. E., Regev, A. & Hacohen, N. „A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection." *Cell* Dec. 2009, **139**(7): 1255–67.

[182] Sidorenko, Y. & Reichl, U. „Structured Model of Influenza Virus Replication in MDCK Cells". *Biotechnology and Bioengineering* 2004, **88**: 1–14.

[183] Sikora, D., Rocheleau, L., Brown, E. G. & Pelchat, M. „Deep sequencing reveals the eight facets of the influenza A/HongKong/1/1968 (H3N2) virus cap-snatching process“. *Scientific Reports* Aug. 2014, **4**: 6181.

[184] Simon, R. & Roychowdhury, S. „Implementing personalized cancer genomics in clinical trials“. *Nature Reviews Drug Discovery* Apr. 2013, **12**(5): 358–369.

[185] Smola, A. J. & Schölkopf, B. „A Tutorial on Support Vector Regression“. *Statistics and Computing* 2004, **14**(3): 199–222.

[186] Somogyi, E. T., Bouteiller, J.-M., Glazier, J. A., König, M., Medley, J. K., Swat, M. H. & Sauro, H. M. „libRoadRunner: a high performance SBML simulation and analysis library: Table 1.“ *Bioinformatics* June 2015, **31**(20): 3315–3321.

[187] Stanford, N. J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P. & Liebermeister, W. „Systematic Construction of Kinetic Models from Genome-Scale Metabolic Networks“. *PLoS ONE* Nov. 2013, **8**(11). Ed. by Vera, J.: e79195.

[188] Stewart, B. & Wild, C. *World Cancer Report 2014*. International Agency for Research on Cancer. IARC Press, 2014.

[189] Swameye, I., Muller, T. G., Timmer, J., Sandra, O. & Klingmuller, U. „Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling“. *Proceedings of the National Academy of Sciences* Jan. 2003, **100**(3): 1028–1033.

[190] Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., Kazer, S. W., Gaillard, A., Kolb, K. E., Villani, A.-C., Johannessen, C. M., Andreev, A. Y., Allen, E. M. V., Bertagnolli, M., Sorger, P. K., Sullivan, R. J., Flaherty, K. T., Frederick, D. T., Jane-Valbuena, J., Yoon, C. H., Rozenblatt-Rosen, O., Shalek, A. K., Regev, A. & Garraway, L. A. „Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq“. *Science* Apr. 2016, **352**(6282): 189–196.

[191] Toh, K. C., Todd, M. J. & Tütüncü, R. H. „SDPT3 — A Matlab software package for semidefinite programming, Version 1.3“. *Optimization Methods and Software* Jan. 1999, **11**(1-4): 545–581.

[192] Tonia, T., Mettler, A., Robert, N., Schwarzer, G., Seidenfeld, J., Weingart, O., Hyde, C., Engert, A. & Bohlius, J. „Erythropoietin or darbepoetin for patients with cancer.“ *The Cochrane database of systematic reviews* Dec. 2012, **12**: CD003407.

[193] Towbin, H., Staehelin, T. & Gordon, J. „Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications.“ *Proceedings of the National Academy of Sciences* Sept. 1979, **76**(9): 4350–4354.

[194] Trapnell, C., Pachter, L. & Salzberg, S. L. „TopHat: discovering splice junctions with RNA-Seq“. *Bioinformatics* Mar. 2009, **25**(9): 1105–1111.

[195] Tummler, K., Lubitz, T., **Schelker, M.** & Klipp, E. „New types of experimental data shape the use of enzyme kinetics for dynamic network modeling“. *FEBS Journal* Nov. 2013, **281**(2): 549–571.

[196] Van Der Maaten, L. J. P. & Hinton, G. E. „Visualizing high-dimensional data using t-SNE“. *Journal of Machine Learning Research* 2008, **9**: 2579–2605.

[197] Vanlier, J., Tiemann, C. A., Hilbers, P. A. J. & Riel, N. A. W. van. „An integrated strategy for prediction uncertainty analysis“. *Bioinformatics* Feb. 2012, **28**(8): 1130–1135.

[198] Vasilyev, F. F., Lopatnikova, J. A. & Sennikov, S. V. „Optimized flow cytometry protocol for analysis of surface expression of interleukin-1 receptor types I and II“. *Cytotechnology* Feb. 2013, **65**(5): 795–802.

[199] Visser, K. E. D., Eichten, A. & Coussens, L. M. „Paradoxical roles of the immune system during cancer development“. *Nature reviews. Cancer* 2006, **6**(January): 24–37.

[200] Vries, E. de, Tscherne, D. M., Wienholts, M. J., Cobos-Jiménez, V., Scholte, F., García-Sastre, A., Rottier, P. J. M. & Haan, C. A. M. de. „Dissection of the Influenza A Virus Endocytic Routes Reveals Macropinocytosis as an Alternative Entry Pathway“. *PLoS Pathog* Mar. 2011, **7**(3). Ed. by Pekosz, A.: e1001329.

[201] Vuong, Q. H. „Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses“. *Econometrica* Mar. 1989, **57**(2): 307.

[202] Wagner, G. P., Kin, K. & Lynch, V. J. „Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples“. *Theory Biosci.* 2012, **131**(4): 281–285.

[203] Walter, E. *Identifiability of Parametric Models*. Pergamon Press, 1987.

[204] Walter, É. & Pronzato, L. *Identification of parametric models from experimental data*. Communications and control engineering. Berlin, Heidelberg, New York: Springer, 1997.

[205] Wang, L., Wang, S. & Li, W. „RSeQC: quality control of RNA-seq experiments“. *Bioinformatics* June 2012, **28**(16): 2184–2185.

[206] Wang, R.-S., Maron, B. A. & Loscalzo, J. „Systems medicine: evolution of systems biology from bench to bedside“. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* Apr. 2015, **7**(4): 141–161.

[207] Wang, Z., Gerstein, M. & Snyder, M. „RNA-Seq: a revolutionary tool for transcriptomics.“ *Nature Reviews. Genetics* Jan. 2009, **10**(1): 57–63.

[208] Watanabe, T., Watanabe, S. & Kawaoka, Y. „Cellular networks involved in the influenza virus life cycle.“ *Cell host & microbe* June 2010, **7**(6): 427–39.

[209] Weemen, B. V. & Schuurs, A. „Immunoassay using antigen-enzyme conjugates“. *FEBS Letters* June 1971, **15**(3): 232–236.

[210] Weinstein, J., Collisson, E., Mills, G., Shaw, K., Ozenberger, B., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. „The Cancer Genome Atlas Pan-Cancer analysis project.“ *Nature Genetics* Oct. 2013, **45**: 1113–20.

[211] Wilkinson, D. J. „Stochastic modelling for quantitative description of heterogeneous biological systems“. *Nature Reviews Genetics* Feb. 2009, **10**(2): 122–133.

[212] Wils, S. & De Schutter, E. „STEPS: Modeling and Simulating Complex Reaction-Diffusion Systems with Python.“ *Front. Neuroinform.* Jan. 2009, **3**(June): 15.

[213] Wu, H., Liu, X., Jaenisch, R. & Lodish, H. F. „Generation of committed erythroid BFU-E and CFU-E progenitors does not require erythropoietin or the erythropoietin receptor“. *Cell* Oct. 1995**,** **83**(1): 59–67.

[214] Yamanaka, K., Ishihama, A. & Nagata, K. „Reconstitution of influenza virus RNA-nucleoprotein complexes structurally resembling native viral ribonucleoprotein cores“. *Journal of Biological Chemistry* 1990**,** **265**(19): 11151–11155.

[215] Yang, P.-C. & Mahmood, T. „Western blot: Technique, theory, and trouble shooting“. *North American Journal of Medical Sciences* 2012**,** **4**(9): 429.

[216] Zaraket, H., Bridges, O. a. & Russell, C. J. „The pH of activation of the hemagglutinin protein regulates H5N1 influenza virus replication and pathogenesis in mice.“ *Journal of virology* May 2013**,** **87**(9): 4826–34.

[217] Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. & Bielas, J. H. „Massively parallel digital transcriptional profiling of single cells“. *Nature Communications* Jan. 2017**,** **8**: 14049.

[218] Zi, Z. & Klipp, E. „Constraint-Based Modeling and Kinetic Analysis of the Smad Dependent TGF-β Signaling Pathway“. *PLoS ONE* Sept. 2007**,** **2**(9). Ed. by Stolovitzky, G.: e936.

[219] Zi, Z. & Klipp, E. „SBML-PET: a Systems Biology Markup Language-based parameter estimation tool.“ *Bioinformatics* Nov. 2006**,** **22**(21): 2704–5.

[220] Ziemann, M., Eren, Y. & El-Osta, A. „Gene name errors are widespread in the scientific literature“. *Genome Biology* Aug. 2016**,** **17**(1).

[221] Zola, H., Swart, B., Banham, A., Barry, S., Beare, A., Bensussan, A., Boumsell, L., Buckley, C. D., Bühring, H.-J., Clark, G., Engel, P., Fox, D., Jin, B.-Q., Macardle, P. J., Malavasi, F., Mason, D., Stockinger, H. & Yang, X. „CD molecules 2006 — Human cell differentiation molecules“. *Journal of Immunological Methods* Jan. 2007**,** **319**(1-2): 1–5.

A

Mathematical modeling of the influenza A infection

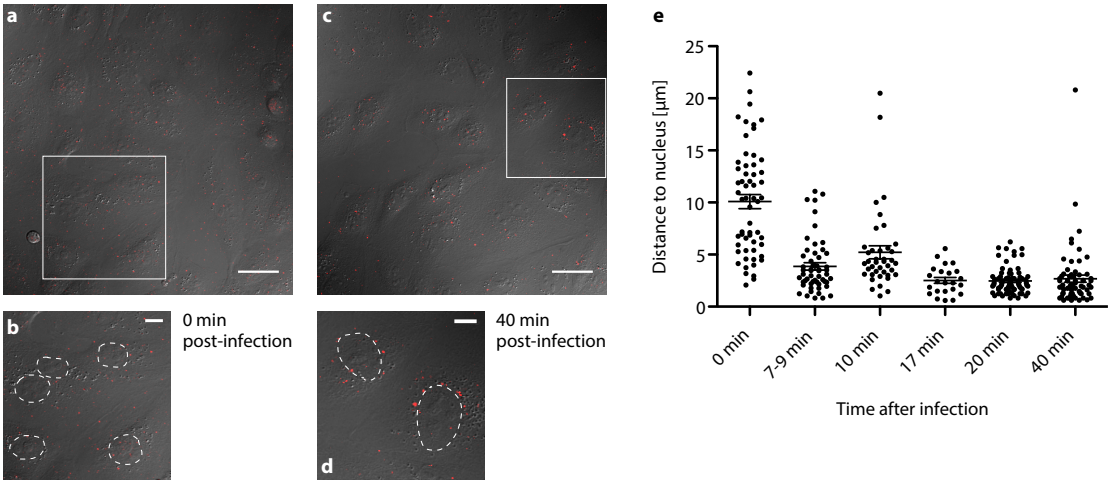

Fig. A.1.: Spatial progression of intracellular influenza A virus-endosome fusion events with respect to the nucleus. MDCK cells were incubated with R18-labeled influenza A X-31 virus for $10\,\mathrm{min}$ at $4\,^{\circ}\mathrm{C}$, washed and R18 was detected using confocal fluorescence microscopy. **(a)** and **(c)** Overview images at 0 and $20\,\mathrm{min}$ post-infection. **(b)** and **(d)** Zoomed representation of the highlighted areas in (a) and (c). Nuclei are marked by a dashed line. Scale bars are $20\,\mu\mathrm{m}$ (a, c) and $5\,\mu\mathrm{m}$ (b, d). **(e)** Fusion events can be detected from R18 dequenching and their shortest distance to the nucleus was measured using ImageJ. Figure adapted from **Schelker** *et al.* [176].

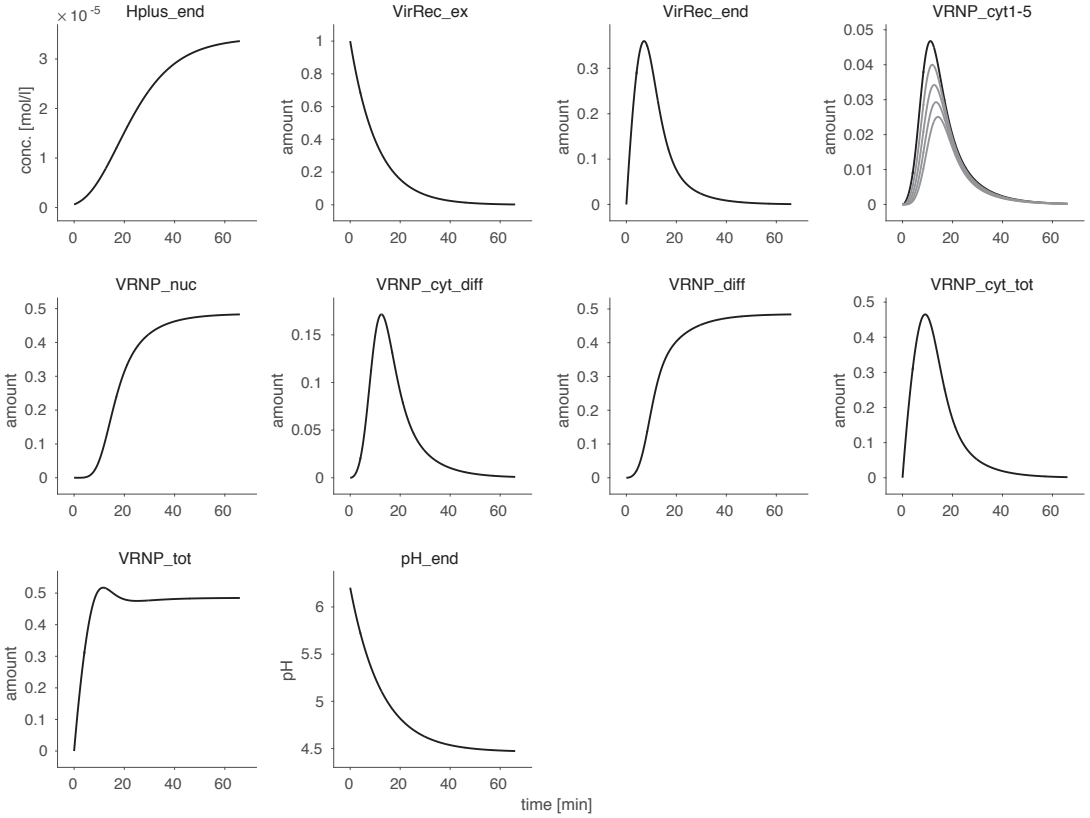

Fig. A.2.: Trajectories of the dynamic IAV entry model. Most variables describe particle numbers as indicated by the label *amount*. Endosomal pH levels are shown as H^+ concentrations an as a pH curve. The grayed lines for the variable $\mathrm{vRNP}_{\mathrm{cyt},1-5}$ denote the pseudo-variables of the linear chain introduced for the delay between vRNP release and nuclear import.

Tab. A.1.: Estimated parameter values. $\hat{\theta}$ indicates the estimated value of the parameters. θ_{lb} and θ_{ub} indicate the lower and upper bounds imposed during optimization. The non-log-column indicates the non-logarithmic value of the estimate. Parameters highlighted in red color indicate parameter values close to their bounds.

index	name	θ_{lb}	$\hat{\theta}$	θ_{ub}	non-log $\hat{\theta}$
1	k_ATPase	-5	-1.1001	+3	$+7.94 \cdot 10^{-02}$
2	h	-5	+0.2796	+3	$+1.90 \cdot 10^{+00}$
3	h_WSN_H3_mut	-5	+0.6285	+3	$+4.25 \cdot 10^{+00}$
4	h_WSN_H3_wt	-5	+0.3599	+3	$+2.29 \cdot 10^{+00}$
5	init_Endosomes_FDQ	-5	-4.8812	+3	$+1.31 \cdot 10^{-05}$
6	init_VirRec_end_140807	-5	-2.1374	+3	$+7.29 \cdot 10^{-03}$
7	init_VirRec_end_140822	-5	-1.1345	+3	$+7.34 \cdot 10^{-02}$
8	init_VirRec_end_150604	-5	-0.5521	+3	$+2.80 \cdot 10^{-01}$
9	init_VirRec_end_150612	-5	-3.6796	+3	$+2.09 \cdot 10^{-04}$
10	init_Virus_number_X31	-5	+0.5787	+3	$+3.79 \cdot 10^{+00}$
11	init_Virus_number_X31_nuclear_NP	-5	+0.7553	+3	$+5.69 \cdot 10^{+00}$
12	init_Virus_number_X31_pH_end	-5	-4.3996	+3	$+3.98 \cdot 10^{-05}$
13	k_Hplus	-7	-5.3133	+3	$+4.86 \cdot 10^{-06}$
14	k_Hplus_WSN_H3_mut	-7	-5.6787	+3	$+2.10 \cdot 10^{-06}$
15	k_Hplus_WSN_H3_wt	-7	-5.3621	+3	$+4.34 \cdot 10^{-06}$
16	k_basal	-5	-2.0345	+3	$+9.24 \cdot 10^{-03}$
17	k_deg	-5	-0.7242	+3	$+1.89 \cdot 10^{-01}$
18	k_end	-5	-1.0307	+3	$+9.32 \cdot 10^{-02}$
19	k_fus	-5	-0.4449	+3	$+3.59 \cdot 10^{-01}$
20	k_imp	-5	+4.0000	+4	$+1.00 \cdot 10^{+04}$
21	k_inhib_100uM	-5	-3.7360	+3	$+1.84 \cdot 10^{-04}$
22	k_inhib_40uM	-5	+3.9953	+4	$+9.89 \cdot 10^{+03}$
23	k_tau	-5	+0.0821	+3	$+1.21 \cdot 10^{+00}$
24	offset_Fus_DiI_DiO	-5	-0.4321	+3	$+3.70 \cdot 10^{-01}$
25	offset_Fus_R18	-5	-1.2652	+3	$+5.43 \cdot 10^{-02}$
26	offset_NP	-5	-0.2488	+3	$+5.64 \cdot 10^{-01}$
27	offset_vRNP_140807	-5	-2.4147	+3	$+3.85 \cdot 10^{-03}$
28	offset_vRNP_140822	-5	-2.4716	+3	$+3.38 \cdot 10^{-03}$
29	offset_vRNP_150604	-5	-3.0353	+3	$+9.22 \cdot 10^{-04}$
30	offset_vRNP_150612	-5	-3.1259	+3	$+7.48 \cdot 10^{-04}$
31	pH_lb	-5	+0.6498	+3	$+4.46 \cdot 10^{+00}$
32	pH_ub	-5	+0.7926	+3	$+6.20 \cdot 10^{+00}$
33	scale_Fus_DiI_DiO	-5	-0.4557	+3	$+3.50 \cdot 10^{-01}$
34	scale_Fus_R18	-5	-0.3111	+3	$+4.89 \cdot 10^{-01}$
35	scale_NP	-5	-0.8445	+3	$+1.43 \cdot 10^{-01}$
36	scale_pH	-5	-0.7445	+3	$+1.80 \cdot 10^{-01}$
37	scale_vRNP_140807	-5	+0.0996	+3	$+1.26 \cdot 10^{+00}$
38	scale_vRNP_140822	-5	-0.8552	+3	$+1.40 \cdot 10^{-01}$
39	scale_vRNP_150604	-5	-2.7923	+3	$+1.61 \cdot 10^{-03}$
40	scale_vRNP_150612	-5	+1.0336	+3	$+1.08 \cdot 10^{+01}$
41	sd_FDQ_H3_mut	-5	-1.2285	+3	$+5.91 \cdot 10^{-02}$
42	sd_FDQ_H3_mut_rel	-5	-1.0824	+3	$+8.27 \cdot 10^{-02}$

43	sd_FDQ_WSN_H3_wt	-5	-1.1098	+3	$+7.77 \cdot 10^{-02}$
44	sd_FDQ_WSN_H3_wt_rel	-5	-1.4075	+3	$+3.91 \cdot 10^{-02}$
45	sd_FDQ_X31	-5	-1.8986	+3	$+1.26 \cdot 10^{-02}$
46	sd_FDQ_X31_rel	-5	-0.8759	+3	$+1.33 \cdot 10^{-01}$
47	sd_Fus_DiI_DiO	-5	-1.5136	+3	$+3.06 \cdot 10^{-02}$
48	sd_Fus_R18	-5	-1.2216	+3	$+6.00 \cdot 10^{-02}$
49	sd_NP	-5	-1.5560	+3	$+2.78 \cdot 10^{-02}$
50	sd_pH_MDCK	-5	-1.5457	+3	$+2.85 \cdot 10^{-02}$
51	sd_vRNP_140807	-5	-1.1650	+3	$+6.84 \cdot 10^{-02}$
52	sd_vRNP_140822	-5	-1.1602	+3	$+6.91 \cdot 10^{-02}$
53	sd_vRNP_150604	-5	-0.6037	+3	$+2.49 \cdot 10^{-01}$
54	sd_vRNP_150612	-5	-1.0648	+3	$+8.61 \cdot 10^{-02}$
55	sd_vRNP_tot_140807	-5	-1.6770	+3	$+2.10 \cdot 10^{-02}$
56	sd_vRNP_tot_140822	-5	-0.3950	+3	$+4.03 \cdot 10^{-01}$
57	sd_vRNP_tot_150604	-5	-0.9325	+3	$+1.17 \cdot 10^{-01}$
58	sd_vRNP_tot_150612	-5	-0.9742	+3	$+1.06 \cdot 10^{-01}$

B Computational approaches for optimized treatment of cancer-associated health conditions

B.1 Optimized treatment strategies for anemia patients based on a mechanistic multi-scale model

B.1.1 Auxiliary ESA-EpoR model

$$\frac{\mathrm{d}[\mathrm{SAv}]}{\mathrm{d}t} = -[\mathrm{SAv}] \cdot [\mathrm{EpoR}] \cdot k_{\mathrm{on\_SAv}} + [\mathrm{SAv\_EpoR}] \cdot k_{\mathrm{off\_SAv}} + [\mathrm{SAv\_EpoR\_i}] \cdot k_{\mathrm{ex}} \tag{B.1}$$

$$\frac{\mathrm{d}[\mathrm{EpoR}]}{\mathrm{d}t} = -[\mathrm{SAv}] \cdot [\mathrm{EpoR}] \cdot k_{\mathrm{on\_SAv}} + [\mathrm{SAv\_EpoR}] \cdot k_{\mathrm{off\_SAv}} + \mathrm{SAv}_{\mathrm{bind}} \cdot k_{\mathrm{t}} - [\mathrm{EpoR}] \cdot k_{\mathrm{t}} + [\mathrm{SAv\_EpoR\_i}] \cdot k_{\mathrm{ex\_SAv}} \tag{B.2}$$

$$\frac{\mathrm{d}[\mathrm{SAv\_EpoR}]}{\mathrm{d}t} = [\mathrm{SAv}] \cdot [\mathrm{EpoR}] \cdot k_{\mathrm{on\_SAv}} - [\mathrm{SAv\_EpoR}] \cdot k_{\mathrm{off\_SAv}} - [\mathrm{SAv\_EpoR}] \cdot k_{\mathrm{e}} \tag{B.3}$$

$$\frac{\mathrm{d}[\mathrm{SAv\_EpoR\_i}]}{\mathrm{d}t} = [\mathrm{SAv\_EpoR}] \cdot k_{\mathrm{e}} - [\mathrm{SAv\_EpoR\_i}] \cdot k_{\mathrm{ex\_SAv}} - [\mathrm{SAv\_EpoR\_i}] \cdot k_{\mathrm{di}} - [\mathrm{SAv\_EpoR\_i}] \cdot k_{\mathrm{de}} \tag{B.4}$$

$$\frac{\mathrm{d}[\mathrm{dSAv\_i}]}{\mathrm{d}t} = [\mathrm{SAv\_EpoR\_i}] \cdot k_{\mathrm{di}} \tag{B.5}$$

$$\frac{\mathrm{d}[\mathrm{dSAv\_e}]}{\mathrm{d}t} = [\mathrm{SAv\_EpoR\_i}] \cdot k_{\mathrm{de}} \tag{B.6}$$

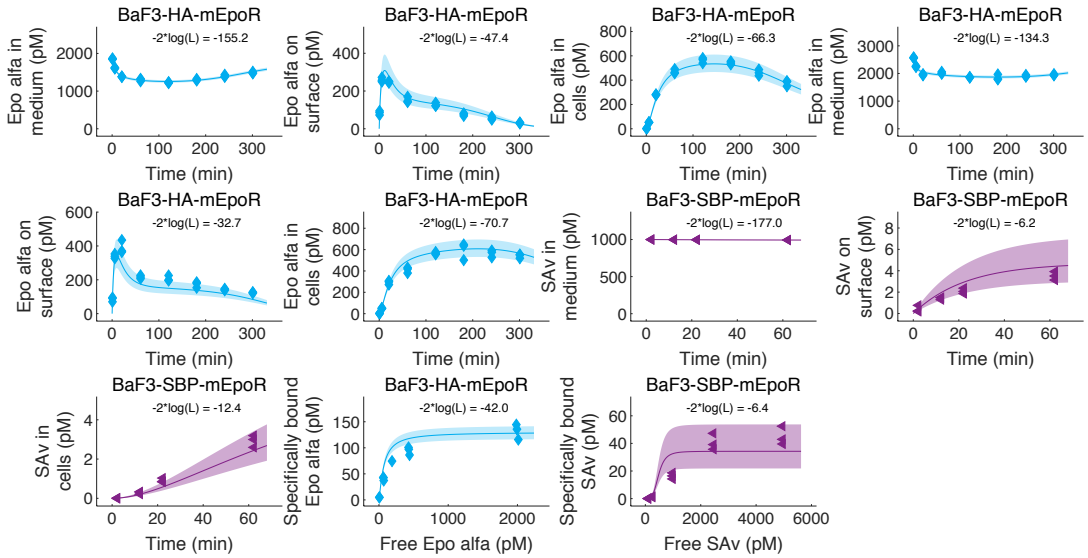

Fig. B.1.: Model simulations and data for the revised model and data by Becker *et al.* [12]. Lines and symbols in light blue represent BaF3-HA-mEpoR cells stimulated with Epo alfa, lines and symbols in purple depict BaF3-SBP-mEpoR cells stimulated with SAv. Shading represents standard deviation of the data as estimated by a parametric error model.

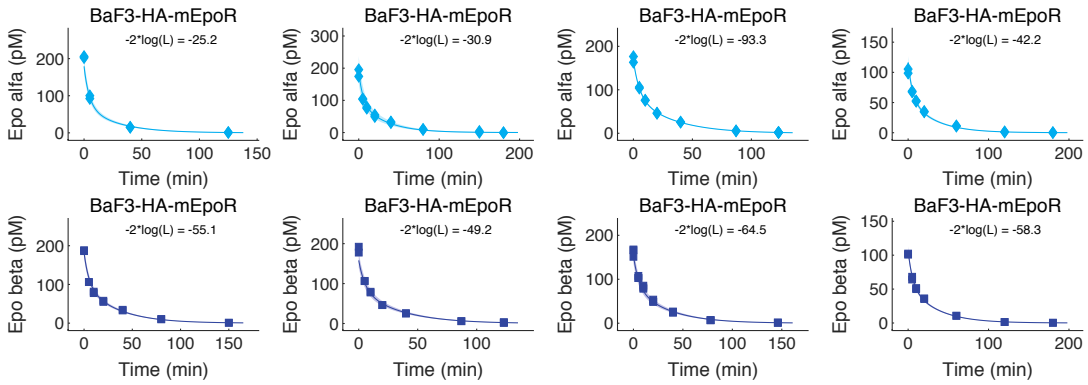

Fig. B.2.: Depletion of ESA in the supernatant of BaF3-HA-mEpoR cells. Colors indicate ESA (Epo alfa, light blue; Epo beta, dark blue). Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model.

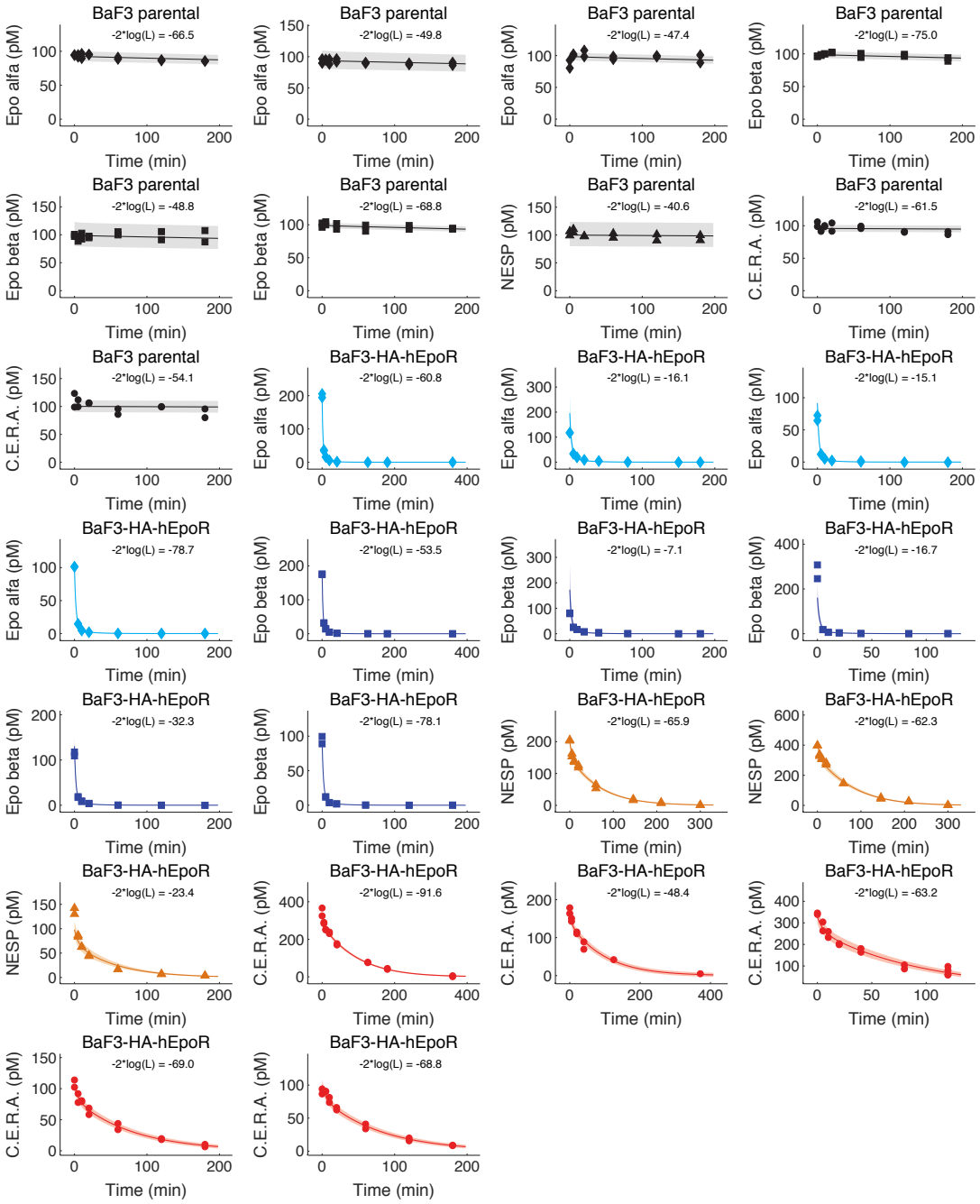

Fig. B.3.: Depletion of ESA in the supernatant of parental BaF3 and BaF3-HA-hEpoR cells. Colors indicate ESA (Epo alfa, light blue; Epo beta, dark blue; NESP, orange; C.E.R.A., red). Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model.

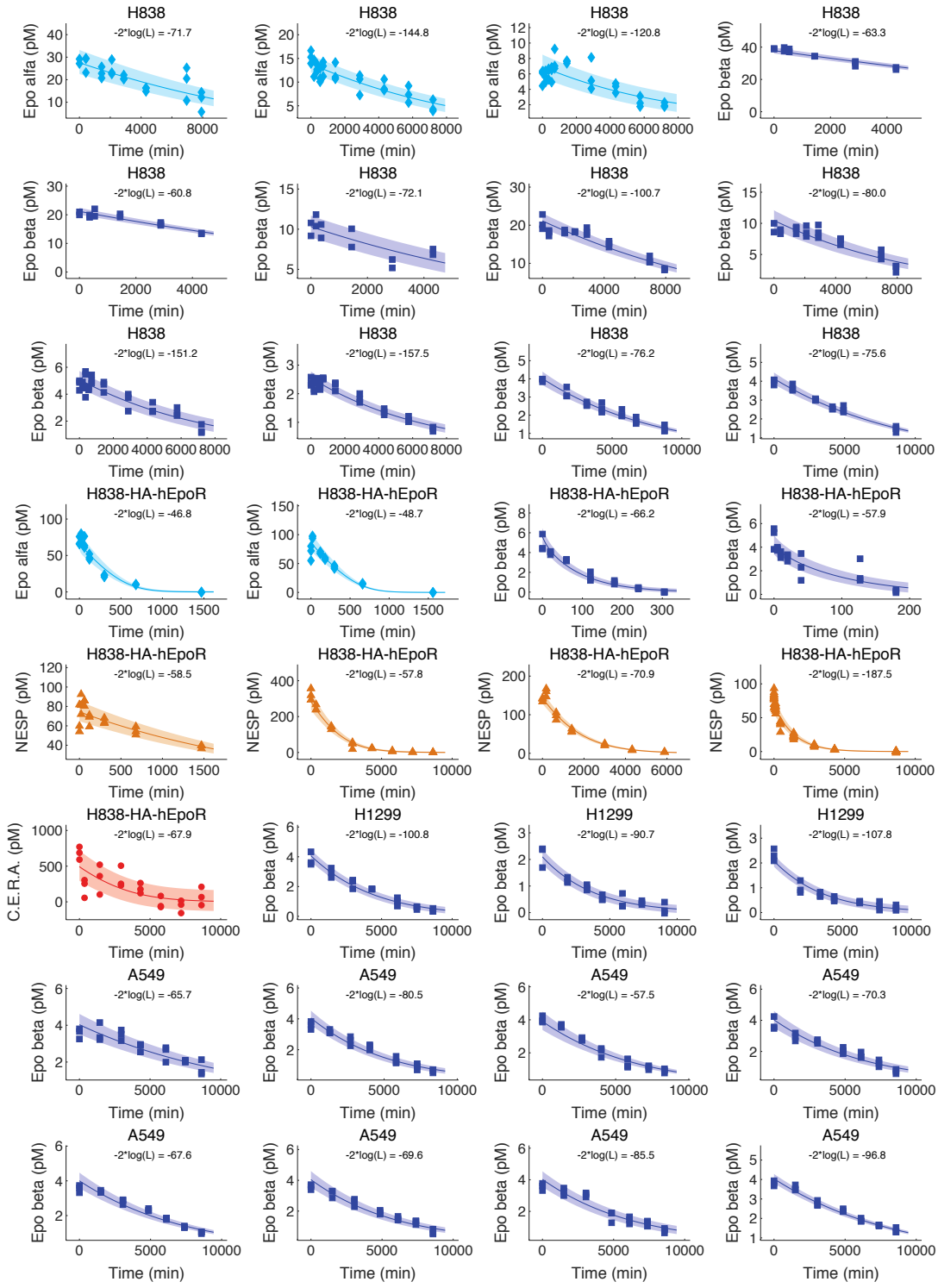

Fig. B.4.: Depletion of ESA in the supernatant of NSCLC cell lines. Colors indicate ESA (Epo alfa, light blue; Epo beta, dark blue; NESP, orange; C.E.R.A., red). Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model.

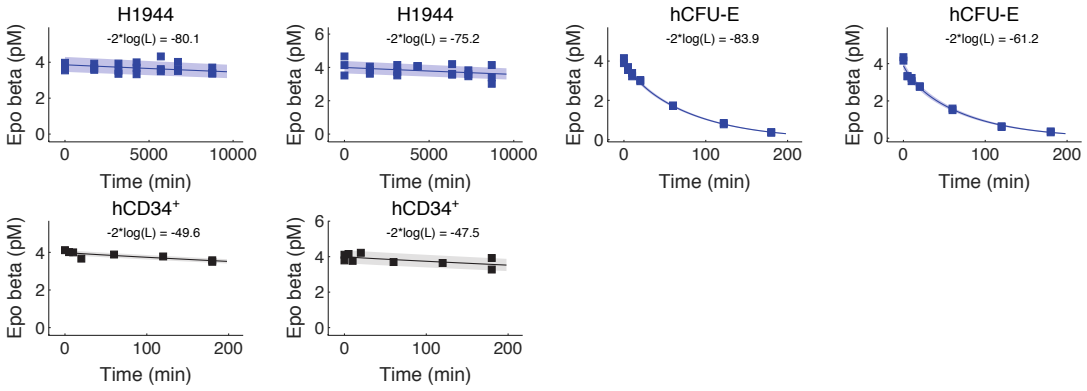

Fig. B.5.: Depletion of ESA in the supernatant of the NSCLC cell line H1944 and of hCFU-E and hCD34$^+$ cells. Colors indicate ESA (Epo alfa, light blue; Epo beta, dark blue; NESP, orange; C.E.R.A., red). Measured data are depicted by symbols, model curves of the core ESA-EpoR model are indicated by lines. Shading represents standard deviation of the data as estimated by a parametric error model.

Tab. B.1.: Parameter values and likelihood-based 68 %-confidence bounds. Asymmetric confidence bounds were obtained from likelihood profiles (Figures B.6 and B.7).

Parameter	$\log_{10}(p)$	$\log_{10}(\text{conf}_{\text{lb}})$	$\log_{10}(\text{conf}_{\text{ub}})$
ESA_bind_A549	−0.112	−0.133	−0.095
ESA_bind_H1299	0.212	0.142	0.271
ESA_bind_H1944	−0.922	−1.180	−0.772
ESA_bind_H838	0.134	0.116	0.163
ESA_bind_H838_HA_hEpoR	2.022	2.012	2.031
ESA_bind_hCFU_E	0.857	0.837	0.880
ESA_bind_hCD34pos	−0.460	−0.653	−0.224
ESA_bind_BaF3_HA_EpoR	2.114	2.092	2.134
ESA_bind_BaF3_parental	−0.318	−0.489	−0.200
ESA_bind_SAv	1.535	1.463	1.623
hill_Epo_alfa	0.006	−0.010	0.023
hill_SAv	0.560	0.497	0.617
kD_CERA_hEpoR	3.035	2.974	3.088
kD_Epo_alfa_hEpoR	1.329	1.286	1.372
kD_Epo_alfa_mEpoR	2.303	2.267	2.339
kD_Epo_alfa_radio	1.913	−1.999	2.285
kD_Epo_alfa_radio_21C	2.153	2.096	2.203
kD_Epo_beta_hEpoR	1.334	1.298	1.370
kD_Epo_beta_mEpoR	2.362	2.339	2.417
kD_NESP_hEpoR	2.823	2.748	2.884
kD_SAv	3.529	3.386	3.665
kD_SAv_21C	2.811	2.718	2.940
kde	−1.884	−1.913	−1.856
kdi	−2.906	−3.096	−2.773
ke	−1.300	−1.313	−1.287
kex	−4.000	−4.000	−3.505
kex_SAv	−6.000	−6.000	−2.814
kon_CERA_hEpoR	−4.090	−4.173	−4.006
kon_Epo_alfa_hEpoR	−3.075	−3.101	−3.048
kon_Epo_alfa_mEpoR	−3.655	−3.687	−3.622
kon_Epo_alfa_radio	−4.023	−4.059	−3.987
kon_Epo_beta_hEpoR	−3.037	−3.061	−3.012
kon_Epo_beta_mEpoR	−3.699	−3.752	−3.636
kon_NESP_hEpoR	−3.901	−4.026	−3.770
kon_SAv	−5.221	−5.334	−5.108
kt	−1.771	−1.797	−1.743

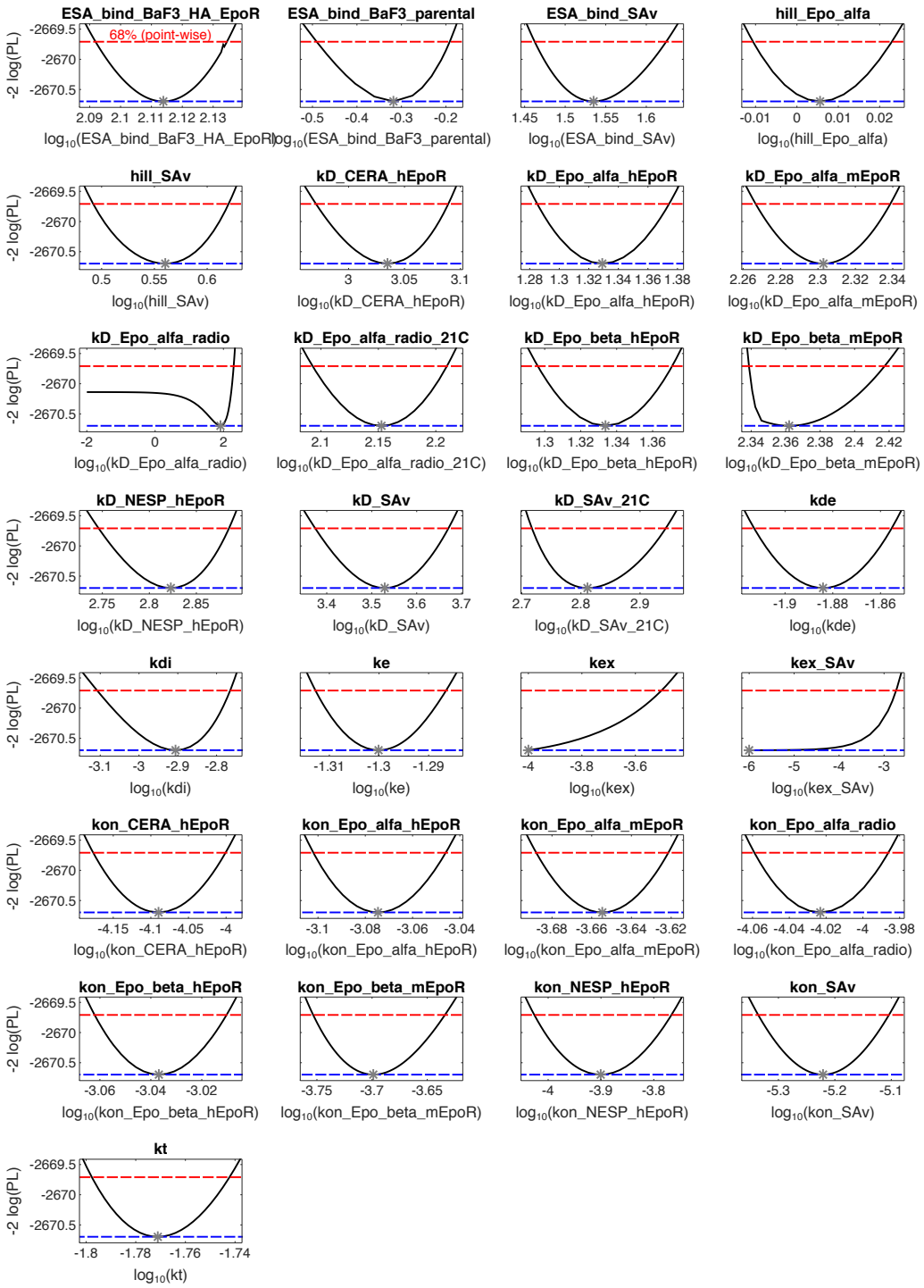

Fig. B.6.: Likelihood profiles for the kinetic parameters of the ESA-EpoR model. Model calibration is based on data from BaF3 cells. The profile likelihood (PL) is calculated as described in Section 2.3.3. The parabola shaped parameter profiles are identifiable. The black solid line denotes the values of the likelihood profile, the gray asterisk indicates the optimal parameter value and the red dashed line represents the 68 %-confidence threshold (point-wise).

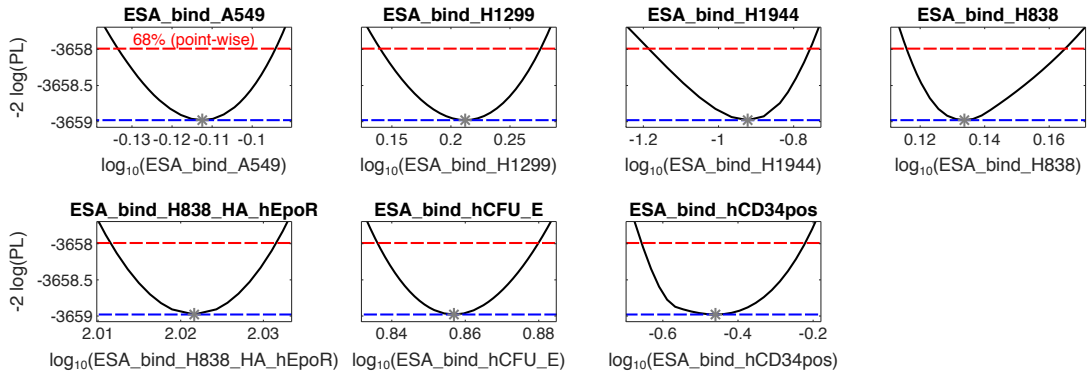

Fig. B.7.: Likelihood profiles for the kinetic parameters of the ESA-EpoR model. Parameters obtained from BaF3 data are fixed to the optimal value given in Table B.1. The profile likelihood (PL) is calculated as described in Section 2.3.3. The parabola shaped parameter profiles are identifiable. The black solid line denotes the values of the likelihood profile, the gray asterisk indicates the optimal parameter value and the red dashed line represents the $68\,\%$-confidence threshold (point-wise).

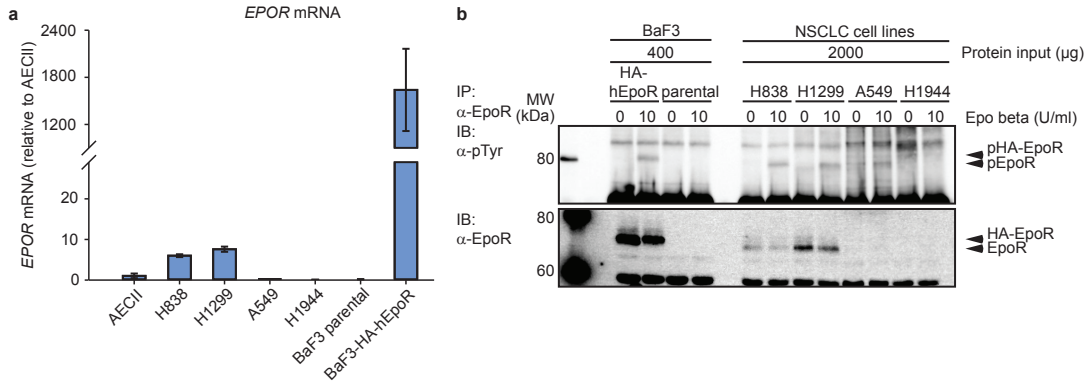

Fig. B.8.: Expression and functionality of EpoR in human lung cancer cell lines. (a) The fold change of *EPOR* mRNA expression in the NSCLC cell lines H838, H1299, A549 and H1944 in comparison to the expression in alveolar epithelial cells type II (AECII), as determined by qRT-PCR, is depicted. The absence of *EPOR* mRNA expression in parental BaF3 cells and the presence of the mRNA due to exogenous expression of the EpoR in BaF3-HA-hEpoR serve as control. The mean of triplicate determinations $\pm$SD is shown. **(b)** Parental BaF3 cells and BaF3-HA-hEpoR as well as the indicated NSCLC cell lines were treated with $10\,\mathrm{U\,ml^{-1}}$ of Epo beta for $10\,\mathrm{min}$ (10) or were left unstimulated (0). Cells were lysed and subjected to immunoprecipitation (IP) with anti-EpoR antibody. Separation of proteins was performed by low-bis $10\,\%$ SDS-PAGE. Detection was performed by immunoblotting (IB) using an anti-phosphotyrosine antibody (pEpoR) or an anti-EpoR antibody (EpoR). As indicated by the protein input for BaF3 cells and NSCLC cell lines ,different amounts of cellular lysates were used for the IP. For EpoR detection, both high and low exposure of the immunoblots are displayed. The position of the pHA-EpoR, pEpoR, HA-EpoR and EpoR are indicated by arrows. MW: molecular weight marker (Magic Marker, Invitrogen). Figure taken from Rodriguez-Gonzalez *et al.* [170].

B.2 Deciphering the cellular composition of unknown patient samples for immunotherapy

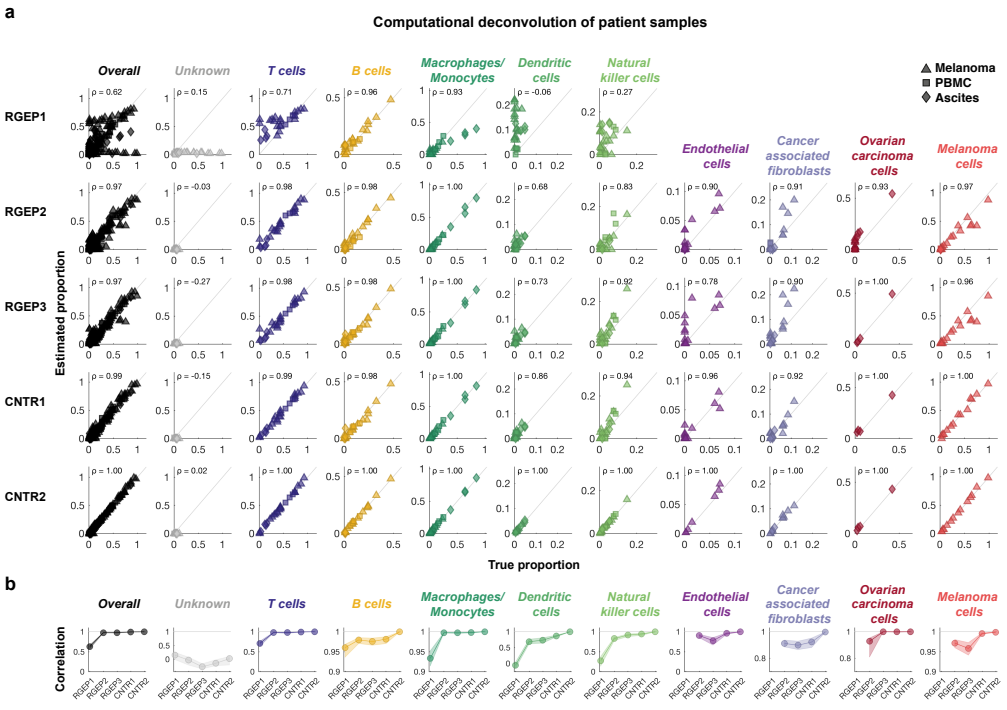

Fig. B.9.: Detailed deconvolution results based on the "mldivide" algorithm and the "Merged" gene set. Caption as in Figure 5.21

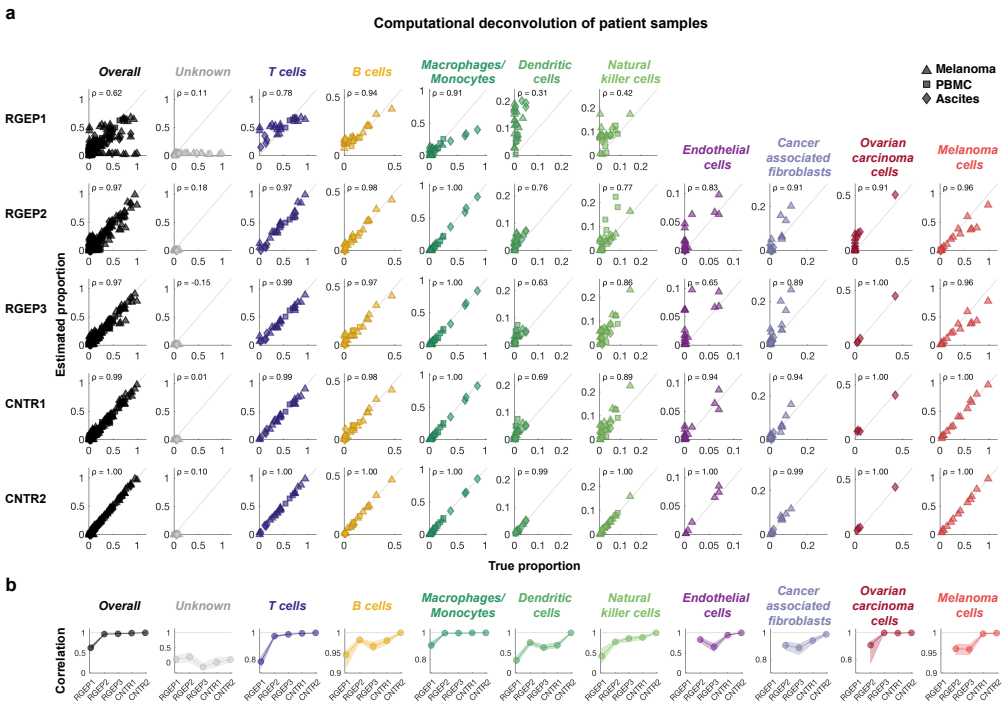

Fig. B.10.: Detailed deconvolution results based on the "SDPT3" algorithm and the "Merged" gene set. Caption as in Figure 5.21

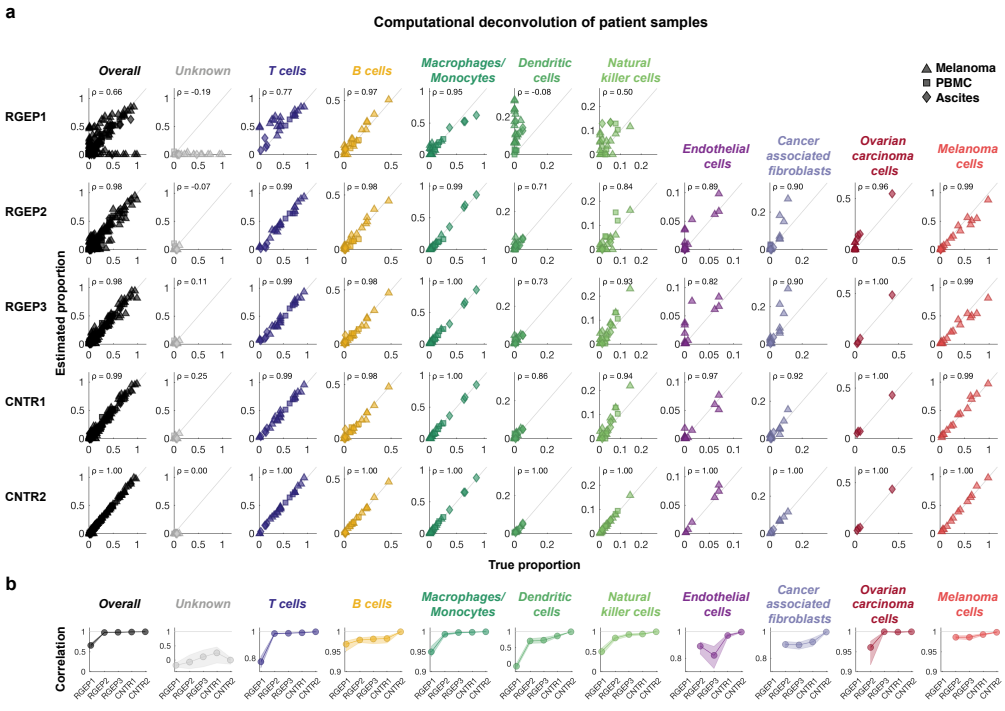

Fig. B.11.: Detailed deconvolution results based on the "fitlm" algorithm and the "Merged" gene set. Caption as in Figure 5.21

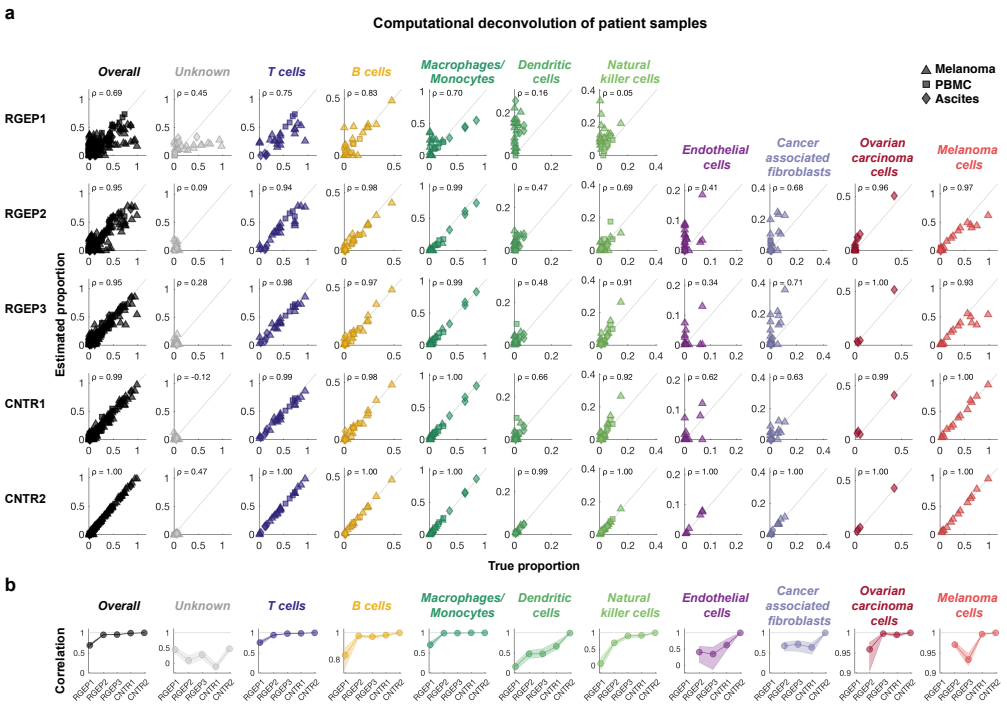

Fig. B.12.: Detailed deconvolution results based on the "Table S12" gene set and the "ν-SVR" algorithm. Caption as in Figure 5.21

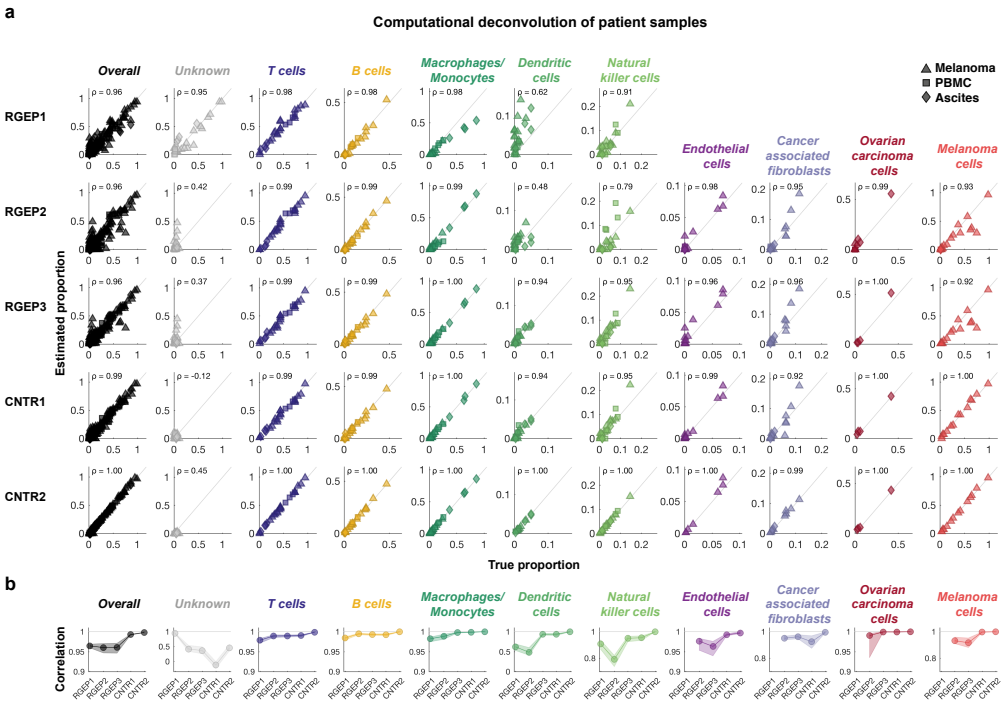

Fig. B.13.: Detailed deconvolution results based on the "Table S3" gene set and the "ν-SVR" algorithm. Caption as in Figure 5.21

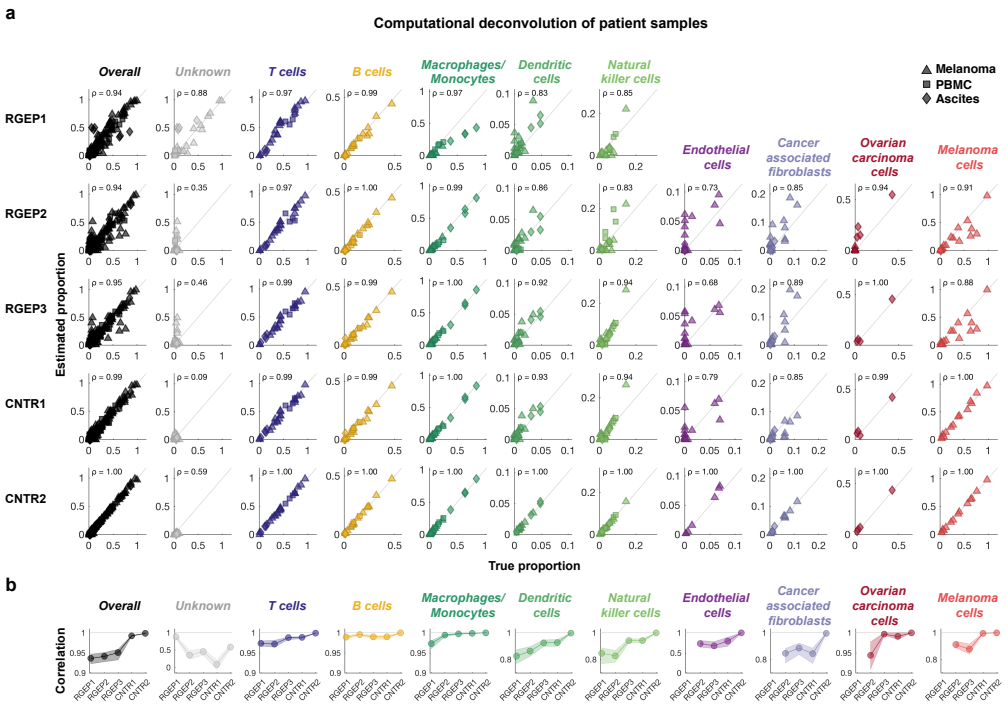

Fig. B.14.: Detailed deconvolution results based on the "LM22" gene set and the "ν-SVR" algorithm. Caption as in Figure 5.21

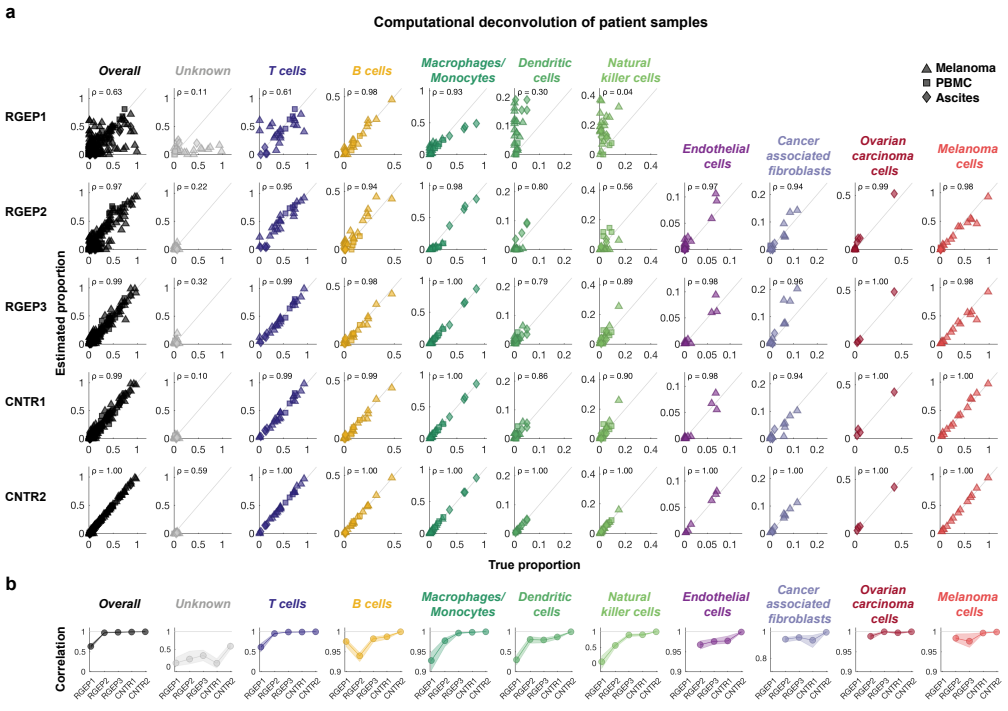

Fig. B.15.: Detailed deconvolution results based on the "All genes" gene set and the "ν-SVR" algorithm. Caption as in Figure 5.21

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Declaration

I hereby certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. No other person's work has been used without due acknowledgment in this thesis. All references have been quoted and all sources of information, including graphs and data sets, have been specifically acknowledged.

Berlin, May 2, 2017

Max Schelker